IJCSI

International Journal of Computer Science Issues

Volume 8, Issue 5, No 3, September 2011
ISSN (Online): 1694-0814

© IJCSI PUBLICATION

www.IJCSI.org

IJCSI proceedings are currently indexed by:

Cornell University Library

Cogprints

Google scholar

.docstoc
find and share professional documents

ScientificCommons

View my documents on Scribd

BASE
Bielefeld Academic Search Engine

SCIRUS
search engine for science

SciRate.com

CiteSeerx beta

dblp.uni-trier.de
Computer Science Bibliography

Q·Sensei BETA

DOAJ DIRECTORY OF OPEN ACCESS JOURNALS

EBSCO HOST

ProQuest

© IJCSI PUBLICATION 2011
www.IJCSI.org

IJCSI Publicity Board 2011

Dr. Borislav D Dimitrov
Department of General Practice, Royal College of Surgeons in Ireland
Dublin, Ireland

Dr. Vishal Goyal
Department of Computer Science, Punjabi University
Patiala, India

Mr. Nehinbe Joshua
University of Essex
Colchester, Essex, UK

Mr. Vassilis Papataxiarhis
Department of Informatics and Telecommunications
National and Kapodistrian University of Athens, Athens, Greece

IJCSI Editorial Board 2011

Dr Tristan Vanrullen
Chief Editor
LPL, Laboratoire Parole et Langage - CNRS - Aix en Provence, France
LABRI, Laboratoire Bordelais de Recherche en Informatique - INRIA - Bordeaux, France
LEEE, Laboratoire d'Esthétique et Expérimentations de l'Espace - Université d'Auvergne, France

Dr Constantino Malagôn
Associate Professor
Nebrija University
Spain

Dr Lamia Fourati Chaari
Associate Professor
Multimedia and Informatics Higher Institute in SFAX
Tunisia

Dr Mokhtar Beldjehem
Professor
Sainte-Anne University
Halifax, NS, Canada

Dr Pascal Chatonnay
Assistant Professor
MaÎtre de Conférences
Laboratoire d'Informatique de l'Université de Franche-Comté
Université de Franche-Comté
France

Dr Karim Mohammed Rezaul
Centre for Applied Internet Research (CAIR)
Glyndwr University
Wrexham, United Kingdom

Dr Yee-Ming Chen
Professor
Department of Industrial Engineering and Management
Yuan Ze University
Taiwan

Dr Gitesh K. Raikundalia
School of Engineering and Science,
Victoria University
Melbourne, Australia

Dr Vishal Goyal
Assistant Professor
Department of Computer Science
Punjabi University
Patiala, India

Dr Dalbir Singh
Faculty of Information Science And Technology
National University of Malaysia
Malaysia

Dr Natarajan Meghanathan
Assistant Professor
REU Program Director
Department of Computer Science
Jackson State University
Jackson, USA

Dr Deepak Laxmi Narasimha
Department of Software Engineering,
Faculty of Computer Science and Information Technology,
University of Malaya,
Kuala Lumpur, Malaysia

Dr. Prabhat K. Mahanti
Professor
Computer Science Department,
University of New Brunswick
Saint John, N.B., E2L 4L5, Canada

Dr Navneet Agrawal
Assistant Professor
Department of ECE,
College of Technology & Engineering,
MPUAT, Udaipur 313001 Rajasthan, India

Dr Panagiotis Michailidis
Division of Computer Science and Mathematics,
University of Western Macedonia,
53100 Florina, Greece

Dr T. V. Prasad
Professor
Department of Computer Science and Engineering,
Lingaya's University
Faridabad, Haryana, India

Dr Saqib Rasool Chaudhry
Wireless Networks and Communication Centre
261 Michael Sterling Building
Brunel University West London, UK, UB8 3PH

Dr Shishir Kumar
Department of Computer Science and Engineering,
Jaypee University of Engineering & Technology
Raghogarh, MP, India

Dr P. K. Suri
Professor
Department of Computer Science & Applications,
Kurukshetra University,
Kurukshetra, India

Dr Paramjeet Singh
Associate Professor
GZS College of Engineering & Technology,
India

Dr Shaveta Rani
Associate Professor
GZS College of Engineering & Technology,
India

Dr. Seema Verma
Associate Professor,
Department Of Electronics,
Banasthali University,
Rajasthan - 304022, India

Dr G. Ganesan
Professor
Department of Mathematics,
Adikavi Nannaya University,
Rajahmundry, A.P, India

Dr A. V. Senthil Kumar
Department of MCA,
Hindusthan College of Arts and Science,
Coimbatore, Tamilnadu, India

Dr Mashiur Rahman
Department of Life and Coordination-Complex Molecular Science,
Institute For Molecular Science, National Institute of Natural Sciences,
Miyodaiji, Okazaki, Japan

Dr Jyoteesh Malhotra
ECE Department,
Guru Nanak Dev University,
Jalandhar, Punjab, India

Dr R. Ponnusamy
Professor
Department of Computer Science & Engineering,
Aarupadai Veedu Institute of Technology,
Vinayaga Missions University, Chennai, Tamilnadu, India

Dr Nittaya Kerdprasop
Associate Professor
School of Computer Engineering,
Suranaree University of Technology, Thailand

Dr Manish Kumar Jindal
Department of Computer Science and Applications,
Panjab University Regional Centre, Muktsar, Punjab, India

Dr Deepak Garg
Computer Science and Engineering Department,
Thapar University, India

Dr P. V. S. Srinivas
Professor
Department of Computer Science and Engineering,
Geethanjali College of Engineering and Technology
Hyderabad, Andhra Pradesh, India

Dr Sara Moein
Computer Engineering Department
Azad University of Najafabad
Iran

Dr Rajender Singh Chhillar
Professor
Department of Computer Science & Applications,
M. D. University, Haryana, India

N. Jaisankar
Assistant Professor
School of Computing Sciences,
VIT University
Vellore, Tamilnadu, India

EDITORIAL

In this fifth edition of 2011, we bring forward issues from various dynamic computer science fields ranging from system performance, computer vision, artificial intelligence, software engineering, multimedia, pattern recognition, information retrieval, databases, security and networking among others.

Considering the growing interest of academics worldwide to publish in IJCSI, we invite universities and institutions to partner with us to further encourage open-access publications.

As always we thank all our reviewers for providing constructive comments on papers sent to them for review. This helps enormously in improving the quality of papers published in this issue.

Google Scholar reported a large amount of cited papers published in IJCSI. We will continue to encourage the readers, authors and reviewers and the computer science scientific community and interested authors to continue citing papers published by the journal.

It was with pleasure and a sense of satisfaction that we announced in mid March 2011 our 2-year Impact Factor which is evaluated at 0.242. For more information about this please see the FAQ section of the journal.

Apart from availability of the full-texts from the journal website, all published papers are deposited in open-access repositories to make access easier and ensure continuous availability of its proceedings free of charge for all researchers.

We are pleased to present IJCSI Volume 8, Issue 5, No 3, September 2011 (IJCSI Vol. 8, Issue 5, No 3). The acceptance rate for this issue is 31.7%.

IJCSI Editorial Board
September 2011 Issue
ISSN (Online): 1694-0814
© IJCSI Publications
www.IJCSI.org

IJCSI Reviewers Committee 2011

- Mr. Markus Schatten, University of Zagreb, Faculty of Organization and Informatics, Croatia
- Mr. Vassilis Papataxiarhis, Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Athens, Greece
- Dr Modestos Stavrakis, University of the Aegean, Greece
- Dr Fadi KHALIL, LAAS -- CNRS Laboratory, France
- Dr Dimitar Trajanov, Faculty of Electrical Engineering and Information technologies, ss. Cyril and Methodius Univesity - Skopje, Macedonia
- Dr Jinping Yuan, College of Information System and Management,National Univ. of Defense Tech., China
- Dr Alexis Lazanas, Ministry of Education, Greece
- Dr Stavroula Mougiakakou, University of Bern, ARTORG Center for Biomedical Engineering Research, Switzerland
- Dr Cyril de Runz, CReSTIC-SIC, IUT de Reims, University of Reims, France
- Mr. Pramodkumar P. Gupta, Dept of Bioinformatics, Dr D Y Patil University, India
- Dr Alireza Fereidunian, School of ECE, University of Tehran, Iran
- Mr. Fred Viezens, Otto-Von-Guericke-University Magdeburg, Germany
- Dr. Richard G. Bush, Lawrence Technological University, United States
- Dr. Ola Osunkoya, Information Security Architect, USA
- Mr. Kotsokostas N.Antonios, TEI Piraeus, Hellas
- Prof Steven Totosy de Zepetnek, U of Halle-Wittenberg & Purdue U & National Sun Yat-sen U, Germany, USA, Taiwan
- Mr. M Arif Siddiqui, Najran University, Saudi Arabia
- Ms. Ilknur Icke, The Graduate Center, City University of New York, USA
- Prof Miroslav Baca, Faculty of Organization and Informatics, University of Zagreb, Croatia
- Dr. Elvia Ruiz Beltrán, Instituto Tecnológico de Aguascalientes, Mexico
- Mr. Moustafa Banbouk, Engineer du Telecom, UAE
- Mr. Kevin P. Monaghan, Wayne State University, Detroit, Michigan, USA
- Ms. Moira Stephens, University of Sydney, Australia
- Ms. Maryam Feily, National Advanced IPv6 Centre of Excellence (NAV6) , Universiti Sains Malaysia (USM), Malaysia
- Dr. Constantine YIALOURIS, Informatics Laboratory Agricultural University of Athens, Greece
- Mrs. Angeles Abella, U. de Montreal, Canada
- Dr. Patrizio Arrigo, CNR ISMAC, italy
- Mr. Anirban Mukhopadhyay, B.P.Poddar Institute of Management & Technology, India
- Mr. Dinesh Kumar, DAV Institute of Engineering & Technology, India
- Mr. Jorge L. Hernandez-Ardieta, INDRA SISTEMAS / University Carlos III of Madrid, Spain
- Mr. AliReza Shahrestani, University of Malaya (UM), National Advanced IPv6 Centre of Excellence (NAv6), Malaysia
- Mr. Blagoj Ristevski, Faculty of Administration and Information Systems Management - Bitola, Republic of Macedonia
- Mr. Mauricio Egidio Cantão, Department of Computer Science / University of São Paulo, Brazil
- Mr. Jules Ruis, Fractal Consultancy, The netherlands
- Mr. Mohammad Iftekhar Husain, University at Buffalo, USA
- Dr. Deepak Laxmi Narasimha, Department of Software Engineering, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia

- Dr. Paola Di Maio, DMEM University of Strathclyde, UK
- Dr. Bhanu Pratap Singh, Institute of Instrumentation Engineering, Kurukshetra University Kurukshetra, India
- Mr. Sana Ullah, Inha University, South Korea
- Mr. Cornelis Pieter Pieters, Condast, The Netherlands
- Dr. Amogh Kavimandan, The MathWorks Inc., USA
- Dr. Zhinan Zhou, Samsung Telecommunications America, USA
- Mr. Alberto de Santos Sierra, Universidad Politécnica de Madrid, Spain
- Dr. Md. Atiqur Rahman Ahad, Department of Applied Physics, Electronics & Communication Engineering (APECE), University of Dhaka, Bangladesh
- Dr. Charalampos Bratsas, Lab of Medical Informatics, Medical Faculty, Aristotle University, Thessaloniki, Greece
- Ms. Alexia Dini Kounoudes, Cyprus University of Technology, Cyprus
- Dr. Jorge A. Ruiz-Vanoye, Universidad Juárez Autónoma de Tabasco, Mexico
- Dr. Alejandro Fuentes Penna, Universidad Popular Autónoma del Estado de Puebla, México
- Dr. Ocotlán Díaz-Parra, Universidad Juárez Autónoma de Tabasco, México
- Mrs. Nantia Iakovidou, Aristotle University of Thessaloniki, Greece
- Mr. Vinay Chopra, DAV Institute of Engineering & Technology, Jalandhar
- Ms. Carmen Lastres, Universidad Politécnica de Madrid - Centre for Smart Environments, Spain
- Dr. Sanja Lazarova-Molnar, United Arab Emirates University, UAE
- Mr. Srikrishna Nudurumati, Imaging & Printing Group R&D Hub, Hewlett-Packard, India
- Dr. Olivier Nocent, CReSTIC/SIC, University of Reims, France
- Mr. Burak Cizmeci, Isik University, Turkey
- Dr. Carlos Jaime Barrios Hernandez, LIG (Laboratory Of Informatics of Grenoble), France
- Mr. Md. Rabiul Islam, Rajshahi university of Engineering & Technology (RUET), Bangladesh
- Dr. LAKHOUA Mohamed Najeh, ISSAT - Laboratory of Analysis and Control of Systems, Tunisia
- Dr. Alessandro Lavacchi, Department of Chemistry - University of Firenze, Italy
- Mr. Mungwe, University of Oldenburg, Germany
- Mr. Somnath Tagore, Dr D Y Patil University, India
- Ms. Xueqin Wang, ATCS, USA
- Dr. Borislav D Dimitrov, Department of General Practice, Royal College of Surgeons in Ireland, Dublin, Ireland
- Dr. Fondjo Fotou Franklin, Langston University, USA
- Dr. Vishal Goyal, Department of Computer Science, Punjabi University, Patiala, India
- Mr. Thomas J. Clancy, ACM, United States
- Dr. Ahmed Nabih Zaki Rashed, Dr. in Electronic Engineering, Faculty of Electronic Engineering, menouf 32951, Electronics and Electrical Communication Engineering Department, Menoufia university, EGYPT, EGYPT
- Dr. Rushed Kanawati, LIPN, France
- Mr. Koteshwar Rao, K G Reddy College Of ENGG.&TECH,CHILKUR, RR DIST.,AP, India
- Mr. M. Nagesh Kumar, Department of Electronics and Communication, J.S.S. research foundation, Mysore University, Mysore-6, India
- Dr. Ibrahim Noha, Grenoble Informatics Laboratory, France
- Mr. Muhammad Yasir Qadri, University of Essex, UK
- Mr. Annadurai .P, KMCPGS, Lawspet, Pondicherry, India, (Aff. Pondicherry Univeristy, India
- Mr. E Munivel , CEDTI (Govt. of India), India
- Dr. Chitra Ganesh Desai, University of Pune, India
- Mr. Syed, Analytical Services & Materials, Inc., USA

- Mrs. Payal N. Raj, Veer South Gujarat University, India
- Mrs. Priti Maheshwary, Maulana Azad National Institute of Technology, Bhopal, India
- Mr. Mahesh Goyani, S.P. University, India, India
- Mr. Vinay Verma, Defence Avionics Research Establishment, DRDO, India
- Dr. George A. Papakostas, Democritus University of Thrace, Greece
- Mr. Abhijit Sanjiv Kulkarni, DARE, DRDO, India
- Mr. Kavi Kumar Khedo, University of Mauritius, Mauritius
- Dr. B. Sivaselvan, Indian Institute of Information Technology, Design & Manufacturing, Kancheepuram, IIT Madras Campus, India
- Dr. Partha Pratim Bhattacharya, Greater Kolkata College of Engineering and Management, West Bengal University of Technology, India
- Mr. Manish Maheshwari, Makhanlal C University of Journalism & Communication, India
- Dr. Siddhartha Kumar Khaitan, Iowa State University, USA
- Dr. Mandhapati Raju, General Motors Inc, USA
- Dr. M.Iqbal Saripan, Universiti Putra Malaysia, Malaysia
- Mr. Ahmad Shukri Mohd Noor, University Malaysia Terengganu, Malaysia
- Mr. Selvakuberan K, TATA Consultancy Services, India
- Dr. Smita Rajpal, Institute of Technology and Management, Gurgaon, India
- Mr. Rakesh Kachroo, Tata Consultancy Services, India
- Mr. Raman Kumar, National Institute of Technology, Jalandhar, Punjab., India
- Mr. Nitesh Sureja, S.P.University, India
- Dr. M. Emre Celebi, Louisiana State University, Shreveport, USA
- Dr. Aung Kyaw Oo, Defence Services Academy, Myanmar
- Mr. Sanjay P. Patel, Sankalchand Patel College of Engineering, Visnagar, Gujarat, India
- Dr. Pascal Fallavollita, Queens University, Canada
- Mr. Jitendra Agrawal, Rajiv Gandhi Technological University, Bhopal, MP, India
- Mr. Ismael Rafael Ponce Medellín, Cenidet (Centro Nacional de Investigación y Desarrollo Tecnológico), Mexico
- Mr. Supheakmungkol SARIN, Waseda University, Japan
- Mr. Shoukat Ullah, Govt. Post Graduate College Bannu, Pakistan
- Dr. Vivian Augustine, Telecom Zimbabwe, Zimbabwe
- Mrs. Mutalli Vatila, Offshore Business Philipines, Philipines
- Mr. Pankaj Kumar, SAMA, India
- Dr. Himanshu Aggarwal, Punjabi University,Patiala, India
- Dr. Vauvert Guillaume, Europages, France
- Prof Yee Ming Chen, Department of Industrial Engineering and Management, Yuan Ze University, Taiwan
- Dr. Constantino Malagón, Nebrija University, Spain
- Prof Kanwalvir Singh Dhindsa, B.B.S.B.Engg.College, Fatehgarh Sahib (Punjab), India
- Mr. Angkoon Phinyomark, Prince of Singkla University, Thailand
- Ms. Nital H. Mistry, Veer Narmad South Gujarat University, Surat, India
- Dr. M.R.Sumalatha, Anna University, India
- Mr. Somesh Kumar Dewangan, Disha Institute of Management and Technology, India
- Mr. Raman Maini, Punjabi University, Patiala(Punjab)-147002, India
- Dr. Abdelkader Outtagarts, Alcatel-Lucent Bell-Labs, France
- Prof Dr. Abdul Wahid, AKG Engg. College, Ghaziabad, India
- Mr. Prabu Mohandas, Anna University/Adhiyamaan College of Engineering, india
- Dr. Manish Kumar Jindal, Panjab University Regional Centre, Muktsar, India

- Prof Mydhili K Nair, M S Ramaiah Institute of Technnology, Bangalore, India
- Dr. C. Suresh Gnana Dhas, VelTech MultiTech Dr.Rangarajan Dr.Sagunthala Engineering College,Chennai,Tamilnadu, India
- Prof Akash Rajak, Krishna Institute of Engineering and Technology, Ghaziabad, India
- Mr. Ajay Kumar Shrivastava, Krishna Institute of Engineering & Technology, Ghaziabad, India
- Mr. Deo Prakash, SMVD University, Kakryal(J&K), India
- Dr. Vu Thanh Nguyen, University of Information Technology HoChiMinh City, VietNam
- Prof Deo Prakash, SMVD University (A Technical University open on I.I.T. Pattern) Kakryal (J&K), India
- Dr. Navneet Agrawal, Dept. of ECE, College of Technology & Engineering, MPUAT, Udaipur 313001 Rajasthan, India
- Mr. Sufal Das, Sikkim Manipal Institute of Technology, India
- Mr. Anil Kumar, Sikkim Manipal Institute of Technology, India
- Dr. B. Prasanalakshmi, King Saud University, Saudi Arabia.
- Dr. K D Verma, S.V. (P.G.) College, Aligarh, India
- Mr. Mohd Nazri Ismail, System and Networking Department, University of Kuala Lumpur (UniKL), Malaysia
- Dr. Nguyen Tuan Dang, University of Information Technology, Vietnam National University Ho Chi Minh city, Vietnam
- Dr. Abdul Aziz, University of Central Punjab, Pakistan
- Dr. P. Vasudeva Reddy, Andhra University, India
- Mrs. Savvas A. Chatzichristofis, Democritus University of Thrace, Greece
- Mr. Marcio Dorn, Federal University of Rio Grande do Sul - UFRGS Institute of Informatics, Brazil
- Mr. Luca Mazzola, University of Lugano, Switzerland
- Mr. Nadeem Mahmood, Department of Computer Science, University of Karachi, Pakistan
- Mr. Hafeez Ullah Amin, Kohat University of Science & Technology, Pakistan
- Dr. Professor Vikram Singh, Ch. Devi Lal University, Sirsa (Haryana), India
- Mr. M. Azath, Calicut/Mets School of Enginerring, India
- Dr. J. Hanumanthappa, DoS in CS, University of Mysore, India
- Dr. Shahanawaj Ahamad, Department of Computer Science, King Saud University, Saudi Arabia
- Dr. K. Duraiswamy, K. S. Rangasamy College of Technology, India
- Prof. Dr Mazlina Esa, Universiti Teknologi Malaysia, Malaysia
- Dr. P. Vasant, Power Control Optimization (Global), Malaysia
- Dr. Taner Tuncer, Firat University, Turkey
- Dr. Norrozila Sulaiman, University Malaysia Pahang, Malaysia
- Prof. S K Gupta, BCET, Guradspur, India
- Dr. Latha Parameswaran, Amrita Vishwa Vidyapeetham, India
- Mr. M. Azath, Anna University, India
- Dr. P. Suresh Varma, Adikavi Nannaya University, India
- Prof. V. N. Kamalesh, JSS Academy of Technical Education, India
- Dr. D Gunaseelan, Ibri College of Technology, Oman
- Mr. Sanjay Kumar Anand, CDAC, India
- Mr. Akshat Verma, CDAC, India
- Mrs. Fazeela Tunnisa, Najran University, Kingdom of Saudi Arabia
- Mr. Hasan Asil, Islamic Azad University Tabriz Branch (Azarshahr), Iran
- Prof. Dr Sajal Kabiraj, Fr. C Rodrigues Institute of Management Studies (Affiliated to University of Mumbai, India), India
- Mr. Syed Fawad Mustafa, GAC Center, Shandong University, China

- Dr. Natarajan Meghanathan, Jackson State University, Jackson, MS, USA
- Prof. Selvakani Kandeeban, Francis Xavier Engineering College, India
- Mr. Tohid Sedghi, Urmia University, Iran
- Dr. S. Sasikumar, PSNA College of Engg and Tech, Dindigul, India
- Dr. Anupam Shukla, Indian Institute of Information Technology and Management Gwalior, India
- Mr. Rahul Kala, Indian Institute of Inforamtion Technology and Management Gwalior, India
- Dr. A V Nikolov, National University of Lesotho, Lesotho
- Mr. Kamal Sarkar, Department of Computer Science and Engineering, Jadavpur University, India
- Dr. Mokhled S. AlTarawneh, Computer Engineering Dept., Faculty of Engineering, Mutah University, Jordan, Jordan
- Prof. Sattar J Aboud, Iraqi Council of Representatives, Iraq-Baghdad
- Dr. Prasant Kumar Pattnaik, Department of CSE, KIST, India
- Dr. Mohammed Amoon, King Saud University, Saudi Arabia
- Dr. Tsvetanka Georgieva, Department of Information Technologies, St. Cyril and St. Methodius University of Veliko Tarnovo, Bulgaria
- Dr. Eva Volna, University of Ostrava, Czech Republic
- Mr. Ujjal Marjit, University of Kalyani, West-Bengal, India
- Dr. Prasant Kumar Pattnaik, KIST,Bhubaneswar,India, India
- Dr. Guezouri Mustapha, Department of Electronics, Faculty of Electrical Engineering, University of Science and Technology (USTO), Oran, Algeria
- Mr. Maniyar Shiraz Ahmed, Najran University, Najran, Saudi Arabia
- Dr. Sreedhar Reddy, JNTU, SSIETW, Hyderabad, India
- Mr. Bala Dhandayuthapani Veerasamy, Mekelle University, Ethiopa
- Mr. Arash Habibi Lashkari, University of Malaya (UM), Malaysia
- Mr. Rajesh Prasad, LDC Institute of Technical Studies, Allahabad, India
- Ms. Habib Izadkhah, Tabriz University, Iran
- Dr. Lokesh Kumar Sharma, Chhattisgarh Swami Vivekanand Technical University Bhilai, India
- Mr. Kuldeep Yadav, IIIT Delhi, India
- Dr. Naoufel Kraiem, Institut Superieur d'Informatique, Tunisia
- Prof. Frank Ortmeier, Otto-von-Guericke-Universitaet Magdeburg, Germany
- Mr. Ashraf Aljammal, USM, Malaysia
- Mrs. Amandeep Kaur, Department of Computer Science, Punjabi University, Patiala, Punjab, India
- Mr. Babak Basharirad, University Technology of Malaysia, Malaysia
- Mr. Avinash singh, Kiet Ghaziabad, India
- Dr. Miguel Vargas-Lombardo, Technological University of Panama, Panama
- Dr. Tuncay Sevindik, Firat University, Turkey
- Ms. Pavai Kandavelu, Anna University Chennai, India
- Mr. Ravish Khichar, Global Institute of Technology, India
- Mr Aos Alaa Zaidan Ansaef, Multimedia University, Cyberjaya, Malaysia
- Dr. Awadhesh Kumar Sharma, Dept. of CSE, MMM Engg College, Gorakhpur-273010, UP, India
- Mr. Qasim Siddique, FUIEMS, Pakistan
- Dr. Le Hoang Thai, University of Science, Vietnam National University - Ho Chi Minh City, Vietnam
- Dr. Saravanan C, NIT, Durgapur, India
- Dr. Vijay Kumar Mago, DAV College, Jalandhar, India
- Dr. Do Van Nhon, University of Information Technology, Vietnam
- Dr. Georgios Kioumourtzis, Researcher, University of Patras, Greece
- Mr. Amol D.Potgantwar, SITRC Nasik, India
- Mr. Lesedi Melton Masisi, Council for Scientific and Industrial Research, South Africa

- Dr. Karthik.S, Department of Computer Science & Engineering, SNS College of Technology, India
- Mr. Nafiz Imtiaz Bin Hamid, Department of Electrical and Electronic Engineering, Islamic University of Technology (IUT), Bangladesh
- Mr. Muhammad Imran Khan, Universiti Teknologi PETRONAS, Malaysia
- Dr. Abdul Kareem M. Radhi, Information Engineering - Nahrin University, Iraq
- Dr. Mohd Nazri Ismail, University of Kuala Lumpur, Malaysia
- Dr. Manuj Darbari, BBDNITM, Institute of Technology, A-649, Indira Nagar, Lucknow 226016, India
- Ms. Izerrouken, INP-IRIT, France
- Mr. Nitin Ashokrao Naik, Dept. of Computer Science, Yeshwant Mahavidyalaya, Nanded, India
- Mr. Nikhil Raj, National Institute of Technology, Kurukshetra, India
- Prof. Maher Ben Jemaa, National School of Engineers of Sfax, Tunisia
- Prof. Rajeshwar Singh, BRCM College of Engineering and Technology, Bahal Bhiwani, Haryana, India
- Mr. Gaurav Kumar, Department of Computer Applications, Chitkara Institute of Engineering and Technology, Rajpura, Punjab, India
- Mr. Ajeet Kumar Pandey, Indian Institute of Technology, Kharagpur, India
- Mr. Rajiv Phougat, IBM Corporation, USA
- Mrs. Aysha V, College of Applied Science Pattuvam affiliated with Kannur University, India
- Dr. Debotosh Bhattacharjee, Department of Computer Science and Engineering, Jadavpur University, Kolkata-700032, India
- Dr. Neelam Srivastava, Institute of engineering & Technology, Lucknow, India
- Prof. Sweta Verma, Galgotia's College of Engineering & Technology, Greater Noida, India
- Mr. Harminder Singh BIndra, MIMIT, INDIA
- Dr. Lokesh Kumar Sharma, Chhattisgarh Swami Vivekanand Technical University, Bhilai, India
- Mr. Tarun Kumar, U.P. Technical University/Radha Govinend Engg. College, India
- Mr. Tirthraj Rai, Jawahar Lal Nehru University, New Delhi, India
- Mr. Akhilesh Tiwari, Madhav Institute of Technology & Science, India
- Mr. Dakshina Ranjan Kisku, Dr. B. C. Roy Engineering College, WBUT, India
- Ms. Anu Suneja, Maharshi Markandeshwar University, Mullana, Haryana, India
- Mr. Munish Kumar Jindal, Punjabi University Regional Centre, Jaito (Faridkot), India
- Dr. Ashraf Bany Mohammed, Management Information Systems Department, Faculty of Administrative and Financial Sciences, Petra University, Jordan
- Mrs. Jyoti Jain, R.G.P.V. Bhopal, India
- Dr. Lamia Chaari, SFAX University, Tunisia
- Mr. Akhter Raza Syed, Department of Computer Science, University of Karachi, Pakistan
- Prof. Khubaib Ahmed Qureshi, Information Technology Department, HIMS, Hamdard University, Pakistan
- Prof. Boubker Sbihi, Ecole des Sciences de L'Information, Morocco
- Dr. S. M. Riazul Islam, Inha University, South Korea
- Prof. Lokhande S.N., S.R.T.M.University, Nanded (MH), India
- Dr. Vijay H Mankar, Dept. of Electronics, Govt. Polytechnic, Nagpur, India
- Dr. M. Sreedhar Reddy, JNTU, Hyderabad, SSIETW, India
- Mr. Ojesanmi Olusegun, Ajayi Crowther University, Oyo, Nigeria
- Ms. Mamta Juneja, RBIEBT, PTU, India
- Prof. Chandra Mohan, John Bosco Engineering College, India
- Mr. Nitin A. Naik, Yeshwant Mahavidyalaya, Nanded, India
- Mr. Sunil Kashibarao Nayak, Bahirji Smarak Mahavidyalaya, Basmathnagar Dist-Hingoli., India
- Prof. Rakesh.L, Vijetha Institute of Technology, Bangalore, India
- Mr B. M. Patil, Indian Institute of Technology, Roorkee, Uttarakhand, India

- Mr. Thipendra Pal Singh, Sharda University, K.P. III, Greater Noida, Uttar Pradesh, India
- Prof. Chandra Mohan, John Bosco Engg College, India
- Mr. Hadi Saboohi, University of Malaya - Faculty of Computer Science and Information Technology, Malaysia
- Dr. R. Baskaran, Anna University, India
- Dr. Wichian Sittiprapaporn, Mahasarakham University College of Music, Thailand
- Mr. Lai Khin Wee, Universiti Teknologi Malaysia, Malaysia
- Dr. Kamaljit I. Lakhtaria, Atmiya Institute of Technology, India
- Mrs. Inderpreet Kaur, PTU, Jalandhar, India
- Mr. Iqbaldeep Kaur, PTU / RBIEBT, India
- Mrs. Vasudha Bahl, Maharaja Agrasen Institute of Technology, Delhi, India
- Prof. Vinay Uttamrao Kale, P.R.M. Institute of Technology & Research, Badnera, Amravati, Maharashtra, India
- Mr. Suhas J Manangi, Microsoft, India
- Ms. Anna Kuzio, Adam Mickiewicz University, School of English, Poland
- Mr. Vikas Singla, Malout Institute of Management & Information Technology, Malout, Punjab, India, India
- Dr. Dalbir Singh, Faculty of Information Science And Technology, National University of Malaysia, Malaysia
- Dr. Saurabh Mukherjee, PIM, Jiwaji University, Gwalior, M.P, India
- Dr. Debojyoti Mitra, Sir Padampat Singhania University, India
- Prof. Rachit Garg, Department of Computer Science, L K College, India
- Dr. Arun Kumar Gupta, M.S. College, Saharanpur, India
- Dr. Todor Todorov, Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
- Mr. Akhter Raza Syed, University of Karachi, Pakistan
- Mrs. Manjula K A, Kannur University, India
- Prof. M. Saleem Babu, Department of Computer Science and Engineering, Vel Tech University, Chennai, India
- Dr. Rajesh Kumar Tiwari, GLA Institute of Technology, India
- Dr. V. Nagarajan, SMVEC, Pondicherry university, India
- Mr. Rakesh Kumar, Indian Institute of Technology Roorkee, India
- Prof. Amit Verma, PTU/RBIEBT, India
- Mr. Sohan Purohit, University of Massachusetts Lowell, USA
- Mr. Anand Kumar, AMC Engineering College, Bangalore, India
- Dr. Samir Abdelrahman, Computer Science Department, Cairo University, Egypt
- Dr. Rama Prasad V Vaddella, Sree Vidyanikethan Engineering College, India
- Prof. Jyoti Prakash Singh, Academy of Technology, India
- Mr. Peyman Taher, Oklahoma State University, USA
- Dr. S Srinivasan, PDM College of Engineering, India
- Mr. Muhammad Zakarya, CIIT, Pakistan
- Mr. Williamjeet Singh, Chitkara Institute of Engineering and Technology, India
- Mr. G.Jeyakumar, Amrita School of Engineering, India
- Mr. Harmunish Taneja, Maharishi Markandeshwar University, Mullana, Ambala, Haryana, India
- Dr. Sin-Ban Ho, Faculty of IT, Multimedia University, Malaysia
- Mrs. Doreen Hephzibah Miriam, Anna University, Chennai, India
- Mrs. Mitu Dhull, GNKITMS Yamuna Nagar Haryana, India
- Dr. D.I. George Amalarethinam, Jamal Mohamed College, Bharathidasan University, India

- Mr. Neetesh Gupta, Technocrats Inst. of Technology, Bhopal, India
- Ms. A. Lavanya, Manipal University, Karnataka, India
- Ms. D. Pravallika, Manipal University, Karnataka, India
- Prof. Vuda Sreenivasarao, St. Mary's college of Engg & Tech, India
- Prof. Ashutosh Kumar Dubey, Assistant Professor, India
- Mr. Ranjit Singh, Apeejay Institute of Management, Jalandhar, India
- Mr. Prasad S.Halgaonkar, MIT, Pune University, India
- Mr. Anand Sharma, MITS, Lakshmangarh, Sikar (Rajasthan), India
- Mr. Amit Kumar, Jaypee University of Engineering and Technology, India
- Prof. Vasavi Bande, Computer Science and Engneering, Hyderabad Institute of Technology and Management, India
- Dr. Jagdish Lal Raheja, Central Electronics Engineering Research Institute, India
- Mr G. Appasami, Dept. of CSE, Dr. Pauls Engineering College, Anna University - Chennai, India
- Mr Vimal Mishra, U.P. Technical Education, Allahabad, India
- Dr. Arti Arya, PES School of Engineering, Bangalore (under VTU, Belgaum, Karnataka), India
- Mr. Pawan Jindal, J.U.E.T. Guna, M.P., India
- Prof. Santhosh.P.Mathew, Saintgits College of Engineering, Kottayam, India
- Dr. P. K. Suri, Department of Computer Science & Applications, Kurukshetra University, Kurukshetra, India
- Dr. Syed Akhter Hossain, Daffodil International University, Bangladesh
- Mr. Nasim Qaisar, Federal Urdu Univetrsity of Arts , Science and Technology, Pakistan
- Mr. Mohit Jain, Maharaja Surajmal Institute of Technology (Affiliated to Guru Gobind Singh Indraprastha University, New Delhi), India
- Dr. Shaveta Rani, GZS College of Engineering & Technology, India
- Dr. Paramjeet Singh, GZS College of Engineering & Technology, India
- Prof. T Venkat Narayana Rao, Department of CSE, Hyderabad Institute of Technology and Management , India
- Mr. Vikas Gupta, CDLM Government Engineering College, Panniwala Mota, India
- Dr Juan José Martínez Castillo, University of Yacambu, Venezuela
- Mr Kunwar S. Vaisla, Department of Computer Science & Engineering, BCT Kumaon Engineering College, India
- Prof. Manpreet Singh, M. M. Engg. College, M. M. University, Haryana, India
- Mr. Syed Imran, University College Cork, Ireland
- Dr. Namfon Assawamekin, University of the Thai Chamber of Commerce, Thailand
- Dr. Shahaboddin Shamshirband, Islamic Azad University, Iran
- Dr. Mohamed Ali Mahjoub, University of Monastir, Tunisia
- Mr. Adis Medic, Infosys ltd, Bosnia and Herzegovina
- Mr Swarup Roy, Department of Information Technology, North Eastern Hill University, Umshing, Shillong 793022, Meghalaya, India
- Mr. Suresh Kallam, East China University of Technology, Nanchang, China
- Dr. Mohammed Ali Hussain, Sai Madhavi Institute of Science & Technology, Rajahmundry, India
- Mr. Vikas Gupta, Adesh Instutute of Engineering & Technology, India
- Dr. Anuraag Awasthi, JV Womens University, Jaipur, India
- Dr. Mathura Prasad Thapliyal, Department of Computer Science, HNB Garhwal University (Centr al University), Srinagar (Garhwal), India
- Mr. Md. Rajibul Islam, Ibnu Sina Institute, University Technology Malaysia, Malaysia
- Mr. Adnan Qureshi, University of Jinan, Shandong, P.R.China, P.R.China
- Dr. Jatinderkumar R. Saini, Narmada College of Computer Application, India

- Mr. Mueen Uddin, Universiti Teknologi Malaysia, Malaysia
- Mr. Manoj Gupta, Apex Institute of Engineering & Technology,Jaipur (Affiliated to Rajasthan Technical University,Rajasthan), Indian
- Mr. S. Albert Alexander, Kongu Engineering College, India
- Dr. Shaidah Jusoh, Zarqa Private University, Jordan
- Dr. Dushmanta Mallick, KMBB College of Engineering and Technology, India
- Mr. Santhosh Krishna B.V, Hindustan University, India
- Dr. Tariq Ahamad Ahanger, Kausar College Of Computer Sciences, India
- Dr. Chi Lin, Dalian University of Technology, China
- Prof. VIJENDRA BABU.D, ECE Department, Aarupadai Veedu Institute of Technology, Vinayaka Missions University, India
- Mr. Raj Gaurang Tiwari, Gautam Budh Technical University, India
- Mrs. Jeysree J, SRM University, India
- Dr. C S Reddy, VIT University, India
- Dr. Amit Wason, Rayat-Bahra Institute of Engineering & Bio-Technology, Kharar, India
- Mr. Yousef Naeemi, Mehr Alborz University, Iran
- Mr. Muhammad Shuaib Qureshi, Iqra National University, Peshawar, Pakistan, Pakistan
- Dr Pranam Paul, Narula Institute of Technology Agarpara. Kolkata: 700109; West Bengal, India
- Dr. G. M. Nasira, Sasurie College of Enginering, (Affliated to Anna University of Technology Coimbatore), India
- Dr. Manasawee Kaenampornpan, Mahasarakham University, Thailand
- Mrs. Iti Mathur, Banasthali University, India
- Mr. Avanish Kumar Singh, RRIMT, NH-24, B.K.T., Lucknow, U.P., India
- Mr. Velayutham Pavanasam, Adhiparasakthi Engineering College, Melmaruvathur, India
- Dr. Panagiotis Michailidis, University of Western Macedonia, Greece
- Mr. Amir Seyed Danesh, University of Malaya, Malaysia
- Dr. Terry Walcott, E-Promag Consultancy Group, United Kingdom
- Mr. Farhat Amine, High Institute of Management of Tunis, Tunisia
- Mr. Ali Waqar Azim, COMSATS Institute of Information Technology, Pakistan
- Mr. Zeeshan Qamar, COMSATS Institute of Information Technology, Pakistan
- Dr. Samsudin Wahab, MARA University of Technology, Malaysia
- Mr. Ashikali M. Hasan, CelNet Security, India
- Dr. Binod Kumar, Lakshmi Narayan College of Tech.(LNCT), India
- Mr. B V A N S S Prabhakar Rao, Dept. of CSE, Miracle Educational Society Group of Institutions, Vizianagaram, India
- Dr. T. Abdul Razak, Associate Professor of Computer Science, Jamal Mohamed College (Affiliated to Bharathidasan University, Tiruchirappalli), Tiruchirappalli-620020, India
- Mr. Aurobindo Ogra, University of Johannesburg, South Africa
- Mr. Essam Halim Houssein, Dept of CS - Faculty of Computers and Informatics, Benha - Egypt
- Mr. Rachit Mohan Garg, Jaypee University of Information Technology, India
- Mr. Kamal Kad, Infosys Technologies, Australia
- Mrs. Aditi Chawla, GNIT Group of Institutes, India
- Dr. Kumardatt Ganrje, Pune University, India
- Mr. Merugu Gopichand, JNTU/BVRIT, India
- Mr. Rakesh Kumar, M.M. University, Mullana,Ambala, India
- Mr. M. Sundar, IBM, India
- Prof. Mayank Singh, J.P. Institute of Engineering & Technology, India
- Dr. Saurabh Pal, VBS Purvanchal University, Jaunpur, India

- Mr. Khaleel Ahmad, S.V.S. University, India
- Mr. Amin Zehtabian, Babol Noshirvani University of Technology / Tetta Electronic Company, Iran
- Mr. Rahul Katarya, Department of Information Technology , Delhi Technological University, India
- Dr. Vincent Ele Asor, University of Port Harcourt, Nigeria
- Ms. Prayas Kad, Capgemini Australia Ltd, Australia
- Mr. Alireza Jolfaei, Faculty and Research Center of Communication and Information Technology, IHU, Iran
- Mr. Nitish Gupta, GGSIPU, India
- Dr. Mohd Lazim Abdullah, University of Malaysia Terengganu, Malaysia
- Mr. Rupesh Nasre., Indian Institute of Science, Bangalore., India.
- Mrs. Dimpi Srivastava, Dept of Computer science, Information Technology and Computer Application, MIET, Meerut, India
- Prof. Santosh Balkrishna Patil, S.S.G.M. College of Engineering, Shegaon, India
- Mr. Mohd Dilshad Ansari, Jaypee University of Information Technology Solan (HP), India
- Mr. Ashwani Kumar, Jaypee University of Information Technology Solan(HP), India
- Dr. Abbas Karimi, Faculty of Engineering, I.A.U. Arak Branch, Iran
- Mr. Fahimuddin.Shaik, AITS, Rajampet, India
- Mr. Vahid Majid Nezhad, Islamic Azad University, Iran
- Ms. C. Divya, Dr G R Damodaran College of Science, Coimbatore-641014, Tamilnadu, India
- Prof. D. P. Sharma, AMU, Ethiopia
- Dr. Sukumar Senthilkumar, School of Mathematical Sciences, Universiti Sains Malaysia, Malaysia
- Mr. Sanjay Bhargava, Banasthali University, Jaipur, Rajasthan, India
- Prof. Rajesh Deshmukh, Shri Shankaracharya Institute of Professional Management & Technology, India
- Mr. Shervan Fekri Ershad, shiraz international university, Iran
- Dr. Vladimir Urosevic, Ministry of Interior, Republic of Serbia
- Mr. Ajit Singh, MDU Rohtak, India

TABLE OF CONTENTS

Document Representation and Clustering with WordNet Based Similarity Rough Set Model

Nguyen Chi Thanh and Koichi Yamada

**Department of Management and Information System Science, Nagaoka University of Technology,
Nagaoka-shi, 940-2188 Japan**

Abstract

Most studies on document clustering till date use Vector Space Model (VSM) to represent documents in the document space, where documents are denoted by a vector in a word vector space. The standard VSM does not take into account the semantic relatedness between terms. Thus, terms with some semantic similarity are dealt with in the same way as terms with no semantic relatedness. Since this unconcern about semantics reduces the quality of clustering results, many studies have proposed various approaches to introduce knowledge of semantic relatedness into VSM model. Those approaches give better results than the standard VSM. However they still have their own issues. We propose a new approach as a combination of two approaches, one of which uses Rough Sets theory and co-occurrence of terms, and the other uses WordNet knowledge to solve these issues. Experiments for its evaluation show advantage of the proposed approach over the others.

Keywords: document clustering, document representation, rough sets, text mining.

1. Introduction

Document clustering is an important text mining technique to generate useful information from text collections such as news articles, research papers, books, digital libraries, e-mail messages, and web pages. Text-based document clustering attempts to group documents into clusters where each cluster might represent a topic that is different from topics of the other clusters.

Document clustering algorithms are divided into two categories in general: partitional clustering and hierarchical clustering. Partitional clustering divides a document collection into groups in a single level, while hierarchical clustering creates a tree structure of documents. There are various document clustering methods proposed in recent years, including hierarchical clustering algorithms using results from a k-way partitional clustering solution [1], spherical k-means [2], bisecting k-means [3], frequent term meaning sequences based method [4], k-means with Harmony Search Optimization [5].

Vector space model is a popular model for document representation in document clustering including the above methods. Documents are represented by vectors of weights, where each weight in a vector denotes importance of a term in the document. In the standard VSM, however, semantic relations between terms are not taken into account. Two terms with a close semantic relation and two other terms with no semantic relation are both treated in the same way. This unconcern about semantics could reduce quality of the clustering result.

There are some approaches proposed to deal with this problem. Tolerance Rough Set model (TRSM) [6] and Similarity Rough Set Model (SRSM) [7] extended the vector space model using Rough Sets theory and co-occurrence of terms. TRSM and SRSM have been successfully applied to document clustering. However, the results showed that SRSM had better performance than TRSM and some other conventional methods [7].

There are other approaches that employ WordNet based semantic similarity to enhance the performance of document clustering [8, 9]. They modified the VSM model by readjusting term weights in the document vectors based on its relationships with other terms co-occurring in the document.

SRSM and WordNet based methods performed better results than the standard VSM. However, they still have their own issues as discussed later. We propose a new method by combining their strength and reducing their weakness. The new method uses both Rough Sets theory and WordNet based semantic similarity to define a new representation model of documents. Experimental results show that it gives better clustering results than the other methods discussed in the paper.

The paper is organized by six sections. In Section 2 and Section 3 we discuss SRSM and WordNet semantic similarity based methods, respectively. Section 4 describes our proposed method. Section 5 presents the results of our experiments on document collections. Finally, Section 6 concludes with a summary and discussion about future research.

2. Similarity rough set model

Similarity Rough Set Model is a mathematical model extended from Pawlak's Rough Set model [10] using similarity relation instead of equivalence relation [7]. It is also an expansion from Tolerance Rough Set Model [6] with a tolerance relation.

Equivalence, tolerance and similarity relations are binary relations that could be used to represent relations between terms in document clustering. An equivalence relation must satisfy reflexive, symmetric and transitive properties, while a tolerance relation does not have to satisfy transitive one. A similarity relation must be reflexive, but not required to be symmetric and transitive [11, 12].

TRSM based on a tolerance relation was successfully applied to information retrieval and document clustering in [6, 13, 14]. Recently, SRSM based on a similarity relation was proposed and applied to document clustering by authors of this paper [7]. It showed that SRSM produces better results than TRSM both in quality and robustness, where co-occurrence of terms was used to obtain tolerance and similarity relations, respectively.

SRSM could be defined as follows: Let the pair *apr* = (*U*, *R*) be an approximation space, where *U* is the universe, and $R \subset U$ x U is a similarity relation on *U*.

$r(x)$: $U \rightarrow 2^U$ is an uncertainty function which corresponds to the similarity relation *R* understood as $yRx \Leftrightarrow y \in r(x)$, which might represent that *y* is similar to *x*. $r(x)$ is a similarity class of all objects that are considered to have similar information to *x*. The function $r(x)$ satisfies reflexive property: $x \in r(x)$, however it is not necessary symmetric and transitive.

Given an arbitrary set $X \subset U$, *X* can be characterized by a pair of lower and upper approximations as follows:

$$\underline{apr}(X) = \{x \in U \mid r^{-1}(x) \subset X\}, \qquad (1)$$

$$\overline{apr}(X) = \bigcup_{x \in X} r(x), \qquad (2)$$

where $r^{-1}(x)$ denotes the inverse relation of *R*, which is the class of referent objects to which *x* is similar:

$$r^{-1}(x) = \{y \in U \mid xRy\} \qquad (3)$$

We proposed a new model of document representation for document clustering using the above generalized rough set theory – Similarity Rough Set Model [7]. The new model is defined as follows.

The universe *U* of the approximation space (*U*, *R*) is the set of all terms *T* used in the document vectors. The binary relation *R* is defined by

$$t_j R t_i \Leftrightarrow f_D(t_i, t_j) \geq \alpha . f_D(t_i), \qquad (4)$$

where $f_D(t_i, t_j)$ is the number of documents in the document set *D* in which term t_i and t_j co-occur, $f_D(t_i)$ is the number of documents in *D* in which term t_i occurs and α is a parameter ($0 < \alpha < 1$). The relation *R* defined above is a similarity relation that satisfies only reflexivity.

An uncertainty function $I_\alpha(t_i)$ corresponding to the similarity relation is defined as

$$I_\alpha(t_i) = \{t_j \in U \mid t_j R t_i\}, \qquad (5)$$

where $I_\alpha(t_i)$ is a set of all terms similar to t_i.

The lower and upper approximation of any subset $X \subset T$ based on this model can be obtained using equations (1) and (2), where *U* and *r* are replaced by *T* and I_α , respectively.

In this case, $I_\alpha{}^{-1}(t_i)$ is the set of terms to which t_i is similar, and is defined as

$$I_\alpha{}^{-1}(t_i) = \{t_j \mid f_D(t_i, t_j) \geq \alpha . f_D(t_j)\} \qquad (6)$$

In the document clustering with SRSM (referred to as SRSM later, while the ordinary approach is referred to as VSM), we applied spherical k-means algorithm [2] to term vectors that consists of terms in upper approximations of ordinary document vectors (term sets). The usage of upper approximation could give us better clustering results, because two documents become similar to each other, if one contains many terms similar (in the sense of eq. (4)) to terms in the other even if the two documents do not have many common terms. Since there are many synonyms in natural language in general and people use different terms to represent a certain thing, the upper approximation would give a positive effect on document clustering.

There would be another advantage of using the upper approximation. The number of terms in a document is usually relatively small in comparison with the number of terms in a corpus. Therefore, the document vectors are usually high dimensional and sparse. Hence, document similarity measurements often yield zero values, which can lead to the poor clustering results. Since the proposed approach puts additional terms into document vectors without increasing the dimension, the unwelcome tendency might be mitigated to some extent.

We use tf×idf weighting scheme to calculate the weights of terms in upper approximations of the document vectors. The term weighting method is extended to define weights of terms that are not contained in documents but in the upper approximations. It ensures that such terms have a weight smaller than the weight of any other term in the document. The weight a_{ij} of term t_i in the upper approximation of document d_j is then defined as follows.

$$a_{ij} = \begin{cases} f_{ij} \times \log\left(\dfrac{N}{f_{D(t_i)}}\right) & \text{if } t_i \in d_j \\ \min_{t_h \in d_j} w_{hj} \times \dfrac{\log\left(\dfrac{N}{f_{D(t_i)}}\right)}{1+\log\left(\dfrac{N}{f_{D(t_i)}}\right)} & \text{if } t_i \in \overline{apr}(d_j) \setminus d_j \\ 0 & \text{if } t_i \notin \overline{apr}(d_j) \end{cases} \qquad (7)$$

where f_{ij} is the frequency of term *i* in document *j*, *N* is number of documents, d_j is a set of terms appearing in document *j*, t_h is the term with the smallest weight in the document *j* and w_{hj} is the original weight of term t_h in the document *j*. Then normalization is applied to the upper approximations of document vectors. The cosine similarity measure is used to calculate the similarity between two vectors.

The algorithm is described as follows [7]:

1. Preprocessing (word stemming, stopwords removal).
2. Create document vectors.
    2.a. Obtain sets of terms appearing in documents.
    2.b. Create document vectors using tf×idf.
    2.b. Generate similarity classes of terms based on their co-occurrences.
    2.c. Create vectors of upper approximations of documents using equation (7) and then the vectors are normalized.
3. Apply the clustering algorithm
    3.a. Start with a random partitioning of the vectors of upper approximations of documents, namely $C^{(0)} = \{C_1^{(0)}, C_2^{(0)}, ..., C_k^{(0)}\}$. Let $c_1^{(0)}, c_2^{(0)}, ..., c_k^{(0)}$ denote the centroids of the given partitioning with the index of iteration $t = 0$.
    3.b. For each document vector x_i, $1 \le i \le N$, find the centroid closest in cosine similarity to its upper approximation $\overline{apr}(x_i)$. Then, compute the new partitioning $C^{(t+1)}$ based on the old centroids $c_1^{(t)}$, $c_2^{(t)}, ..., c_k^{(t)}$:
    $C_j^{(t+1)}$ is the set of all document vectors whose upper approximations are closest to the centroid $c_j^{(t)}$. If the upper approximation of a document is closest to more than one centroid, then it is randomly assigned to one of the clusters.
    3.c. Compute the new centroids:

$$s_j = \sum_i \overline{apr}(x_i)\,,\ c_j^{(t+1)} = \frac{s_j}{\|s_j\|}\,,\ 1 \le j \le k\,,$$

    where $c_j^{(t+1)}$ denotes the centroid or the mean of the upper approximations of documents in cluster $C_j^{(t+1)}$.
    3.d. If some "stopping criterion" is met, then set $C_j^* = C_j^{(t+1)}$ and set $c_j^* = c_j^{(t+1)}$ for $1 \le j \le k$, and exit. Otherwise, increment *t* by 1, and go to step 3.b above.

In our implementation, the iteration stops when the centroids of the generated clusters are identical to those generated in the previous iteration.

In SRSM, we used co-occurrence of terms to calculate the semantic relation between terms. The usage of co-occurrence gives us a merit that lets us define similarity relations automatically without any knowledge base. However, it might also have a weakness that in some cases co-occurrence of terms does not necessarily mean they have a similar meaning. In the case, terms that do not appear in a document nor similar to any term in the document may be contained in the upper approximation.

3. WordNet semantic similarity based model

WordNet is an electronic lexical database of English, available to researchers in computational linguistics and natural language processing [15]. WordNet was developed and is being maintained by the Cognitive Science Laboratory of Princeton University. In WordNet, a concept represents a meaning of a term. Terms which have the same concept are grouped in a synset. Each synset has its definition (gloss) and links with other synsets higher or lower in the hierarchy by different types of semantic relations.

There are different methods to compute semantic similarity of terms using WordNet, which can be divided into four categories: path based, information content based, gloss based and vector based methods. Path based methods use length of the path between concept nodes to calculate the similarity relatedness [16, 17]. Information content based methods [18, 19] measure the relatedness of the two concepts using the information content of the most specific shared parent. In gloss based methods [20, 21], glosses of concepts are used to determine the relatedness of concepts. In vector based methods [22, 23], the relatedness between terms are computed using concept vectors derived from glosses.

Recently, some studies used WordNet-based semantic similarity to enhance performance of document clustering [8, 9]. They modified the VSM model by readjusting weights of terms in the documents. The basic idea is that a term is considered more important if other terms semantically related to it appear in the same document. They increase weight values of such terms with the following equation:

$$\tilde{w}_{i_1 j} = w_{i_1 j} + \sum_{\substack{t_{i_2} \in d_j \\ i_2 \ne i_1}} sim(t_{i_1}, t_{i_2}) w_{i_2 j} \qquad (8)$$

where $w_{i_1 j}$ is the original weights of term t_{i_1} in document d_j, $sim(t_{i_1}, t_{i_2})$ is the semantic similarity between the two terms calculated using a WordNet based measure. They proposed improved VSM model based on this idea and showed that the clustering performance based on the new model was better than that based on the VSM.

The advantage of this approach is the high reliability of similarity given by the WordNet. The basic idea behind eq. (8) also seems adequate. A possible weak point might come from the general property of WordNet. Since it is a general dictionary, it might not work for documents in a specific field. Another is that it utilizes the knowledge of similarity only to adjust the importance of terms in a document. It does not let us find similarity between two documents where one contains many terms similar to ones in the other but the two do not have many common terms.

4. WordNet based similarity rough set model for document clustering

In document clustering, the effect of semantic similarity between terms is large, and must be taken into account to enhance the performance of VSM. In SRSM, the semantic relation between terms is calculated using co-occurrence of terms. However, there seem cases when terms have high co-occurrence but have low semantic similarity. WordNet-based approaches measure the relatedness of terms using the lexical database. Based on the ontology structure of terms or definitions of terms in WordNet, we can compute scores of semantic relatedness. However, as a general dictionary, WordNet does not cover all terms and term meanings in every specific subject. Moreover, in different fields, the semantic relation of terms may be different. Our idea is to exploit both approaches to get better clustering results.

In SRSM, we defined the similarity class of terms using the relation R given by eq. (4). Here, we propose a new relation that integrates WordNet knowledge to eliminate terms having no similar meaning but a high frequency of co-occurrence.

$$t_j R t_i \Leftrightarrow f_D(t_i, t_j) \geq \alpha . f_D(t_i) \wedge ((t_i \text{ not in WordNet}) \vee (t_j \text{ not in WordNet}) \vee \text{sim}(t_i, t_j) > \theta)), \quad (9)$$

where θ is a threshold value.

The relation defined by Eq. (9) is a similarity relation, because it is reflexive, non-symmetric and non-transitive. The basic idea is that term t_j is similar to t_i when t_j is similar to t_i from the viewpoint of co-occurrence and they are also similar in the semantics of WordNet. If t_i or t_j is not in WordNet, we use only the co-occurrence similarity.

Then we can define a new representation model based on this relation in the similar way to the one in section 2.

Let the pair $apr = (U, R)$ be an approximation space, where U is the set of all index terms T in the same way as SRSM, and $R \subset U \text{ x } U$ is a similarity relation on U.

$r(x)$: $U \rightarrow 2^U$ is an uncertainty function which corresponds to the relation R understood as $yRx \Leftrightarrow y \in r(x)$, which might represent that y is similar to x. $r(x)$ is a similarity class of all objects that are considered to have similar information to x. The function $r(x)$ satisfies reflexive property: $x \in r(x)$, however it is not necessary symmetric and transitive.

Given an arbitrary set $X \subset U$, X can be characterized by a pair of lower and upper approximations as equations (1) and (2).

The binary relation R is a relation that corresponds to an uncertainty function defined by eq. (9). That is,

$$I_{\alpha\theta}(t_i) = \{t_j \in U \mid t_j R t_i\}. \quad (10)$$

R is a similarity relation because it only satisfies the properties of reflexivity.

In SRSM, we assigned weights to terms that do not occur in the document but belong to similarity classes of terms in the document, and do not change the weight values of terms in the document. In the new method we improve the SRSM by readjusting weight values of terms based on the idea of WordNet based methods.

The weight a_{ij} of term t_i in the upper approximation of document d_j is then defined as follows.

$$a_{ij} = \begin{cases} f_{ij} \times \log\left(\dfrac{N}{f_{D(t_i)}}\right) + \sum\limits_{\substack{t_k \in d_j \\ k \neq i}} sim(t_i, t_k) a_{kj} & \text{if } t_i \in d_j \\ \min_{\mathrm{t_h} \in d_j} a_{hj} \times \dfrac{\log\left(\dfrac{N}{f_{D(t_i)}}\right)}{1 + \log\left(\dfrac{N}{f_{D(t_i)}}\right)} & \text{if } t_i \in \overline{apr}(d_j) \setminus d_j \\ 0 & \text{if } t_i \notin \overline{apr}(d_j) \end{cases} \quad (11)$$

The new proposed approach could be regarded as a combination of SRSM and WSSM (WordNet Semantic Similarity based Model) which incorporate the advantages of both the models. WSSM is completely included in the proposed approach because weights of terms in a document are adjusted using eq. (8). In addition, it is an improved version of SRSM, because it calculates the upper approximation of the term set of a document and uses it as the document vector. The improvements are the similarity relation (eq. (9)) used to calculate the upper approximation, and eq. (11) to readjust the weights of terms that are contained in the document.

5. Experimental results

In the experiments, we use two test collections to evaluate the proposed approach in comparison with SRSM, WSSM and methods in CLUTO toolkit [24]. The algorithms provided in CLUTO toolkit are based on the partitional, agglomerative, and graph-partitioning paradigms. They are denoted as *rb*, *rbr*, *direct*, *agglo*, *graph*, *bagglo*. The *rb* is a repeated bisecting approach. The *rbr* is the same as the repeated bisecting method except that at the end the overall solution is globally optimized. The *direct* is a partition method which uses an iterative refinement algorithm to optimize a global clustering criterion function. The *agglo* is an agglomerative clustering algorithm. The *graph* uses a nearest-neighbor graph to model documents, and then divides the graph into *k* clusters using a min-cut graph partitioning algorithm. In the *bagglo*, agglomeration process is used to cluster documents after the document collection is split into $\sqrt{N}$ clusters using the *rb* method.

The first test collection is a classic data set obtained by combining CACM, CISI, CRANFIELD, and MEDLINE abstracts which is available from [25]. The dataset includes abstracts of papers in different fields. CACM contains 3204 abstracts from Communications of ACM, CISI contains 1460 abstracts of information science papers, CRANDFIELD contains 1400 abstract of aeronautical papers, MEDLINE contains 1033 abstracts of medicine papers. The clustering algorithms are supposed to cluster the dataset containing 7097 abstracts into four groups.

After preprocessing (stemming, stop-words elimination, and high frequency word pruning), we have 13177 terms in the document collection. With the 13177 terms we created 7097 document vectors using tf×idf weighting scheme, each document vector has 13177 dimensions.

We evaluate clustering results obtained by each algorithm with three commonly-used measures: entropy, *F* measure and mutual information [3, 26]. There are different clustering quality measures rendering different results. However, if a method performs better than the others on many of these measures then we could say that the method is better than the others.

Entropy, *F* measure and mutual information measures are external quality measures which evaluate the clustering results by comparing the clusters produced by the algorithm to the known classes of documents. With the entropy measure method, the clustering quality is better if the entropy is smaller. While with *F* measure and mutual information method, the higher the evaluated values are the better clustering result is.

We run the experiments with the proposed method, SRSM and WSSM. We also run the test collection with the CLUTO toolkit.

The WordNet-based similarity measure used in the experiment is the Wu and Palmer measure [17], which is a path-based method. It computes the relatedness of two concepts using the lowest common subsumer of two concepts $lcs(c_1, c_2)$ which is the first shared concept on the paths from the concepts to the root concept of the ontology hierarchy.

$$sim(c_1,c_2)=\frac{2\times depth(lcs)}{l(c_1,lcs)+l(c_2,lcs)+2\times depth(lcs)} \qquad (12)$$

where $l(c_1,lcs)$ is the length of the path between the two nodes and $depth(lcs)$ is the number of nodes on the path from *lcs* to root.

We ran the experiment using the proposed method with the value of threshold $\theta = 0.3$.

Table 1 shows the evaluation of clustering results from CLUTO toolkit's algorithms, Table 2 shows the evaluation of clustering results of SRSM and the newly proposed method with different values of parameter α. The best evaluation in each quality measure is shaded in Table 2. As for WSSM, the evaluation of the clustering result was 0.363, 0.332, 0.894 for entropy, mutual information and *F* measure respectively. As seen in these results, the best case is the clustering by the proposed approach with $\alpha = 0.55$ in all three evaluation measures.

With SRSM, co-occurrence of terms is used to determine the similarity classes of terms. In the proposed method, we use both co-occurrence of terms and WordNet based semantic similarity. The new approach, as suggested by the figures in the column of "size of similarity classes" in Table 2, can remove irrelevant terms from similarity classes. For example, with the SRSM implementation in our experiment, similarity class of “photon” contains “integration” and “algebra”, which have low semantic relatedness with “photon” itself. With the new method, “integration” and “algebra” are removed from the similarity class. Another example would be the term “program”, which for SRSM is in the similarity class of “glossary”, while for the new approach it is not the case. The removal of irrelevant terms improves the quality of similarity classes and could give better clustering results.

Table 1: Clustering results of the first data set from CLUTO toolkit [7]

CLUTO			
Method	Entropy	Mutual information	*F* measure
Rb	0.562	0.261	0.641
Rbr	0.561	0.261	0.651
Direct	0.552	0.264	0.672
Agglo	1.283	0.001	0.452
Bagglo	0.455	0.299	0.712

Table 2: Evaluation of clustering results with SRSM and the new method for the first data set

α	SRSM					New method				
	Entropy	Mutual information	*F* measure	Size of similarity classes		Entropy	Mutual Information	*F* measure	Size of similarity classes	
				Max	Avg				Max	Avg
0.40	0.375	0.328	0.859	83	4.80	0.331	0.344	0.896	75	3.69
0.45	0.348	0.337	0.877	69	3.61	0.319	0.348	0.902	67	2.88
0.50	0.327	0.345	0.892	69	3.52	0.291	0.358	0.915	67	2.81
0.55	0.309	0.352	0.900	60	2.45	0.286	0.360	0.916	60	2.07
0.60	0.309	0.352	0.905	60	2.19	0.288	0.359	0.915	60	1.89
0.65	0.306	0.353	0.907	60	2.15	0.297	0.356	0.913	60	1.86
0.70	0.308	0.353	0.908	28	1.37	0.294	0.358	0.914	20	1.29
0.75	0.311	0.351	0.907	28	1.34	0.299	0.356	0.913	20	1.27
0.80	0.310	0.352	0.908	17	1.21	0.300	0.355	0.913	17	1.17

The maximum, the minimum and the average sizes of similarity classes of SRSM and the proposed method are shown in Table 2. Number of terms that are not in WordNet is 4753 among 13177 terms. It is around 36%.

We can see that sizes of similarity classes of the proposed method are smaller than those of SRSM. The difference is the result of removing terms with low WordNet based semantic relatedness from similarity classes in the proposed method. As defined by eq. (9), a similarity class of a term t_i consists of terms that satisfy both the condition of co-occurrence with t_i and one of the following conditions: 1) at least one of the two terms does not exist in WordNet database; 2) the WordNet based similarity measure between the two terms is greater than a threshold value. For example, when $\alpha = 0.55$, the average number of similarity classes defined only by the co-occurrence condition is 2.45 (SRSM), while the one defined by eq. (9) is 2.07 (the proposed method), which means that 0.38 terms in average are removed from similarity classes of SRSM because they do not satisfy the above condition 1) nor 2). Then, among the remaining 2.07 terms of similarity classes of the proposed method, 0.88 terms satisfy the condition 1) and 1.19 satisfy condition 2), in average.

The contingency table of the best case of the proposed method is shown in Table 3. Precision and recall of CACM, CISI, CRANFIELD, and MEDLINE are 0.964, 0.797, 0.930, 0.962 and 0.851, 0.952, 0.976, 0.983, respectively.

The computation time of the new method is almost same as the one of the SRSM method which has the time complexity of $O(M\log M)$ [7], where M is the number of terms in the text collection. The difference between the new and SRSM methods is the computation of term semantic relationship based on WordNet. The computation of semantic relationship is fast because we use a path based method and the maximum depth of the word hierarchy in WordNet is sixteen [9], a very small number in comparison with number of terms in a text collection.

Table 3: Contingency table of the best case of the proposed method

	CACM	CISI	CRANFIELD	MEDLINE
Cluster 1	2726	68	29	5
Cluster 2	347	1390	4	4
Cluster 3	94	0	1366	9
Cluster 4	37	2	1	1015

The second test collection used in our experiment is abstracts of papers from several IEEE journals of several fields. We formed a collection of 1010 documents from IEEE Transactions on Knowledge and Data Engineering (378 abstracts), IEEE Transactions on Biomedical Engineering (311 abstracts) and IEEE Transactions on Nanotechnology (321 abstracts). These categories of documents are denoted as KDE, BIO and NANO. We use the clustering methods to cluster the data set into three clusters.

After removing stopwords and stemming words, we have 5690 terms in the document collection. With 5690 terms, the algorithm created 1010 document vector using tf×idf weighting scheme, each document vector has 5690 dimensions.

Table 4: Clustering results of the second data set from CLUTO toolkit [7]

	CLUTO		
Method	Entropy	Mutual information	F measure
rb	0.290	0.366	0.898
rbr	0.198	0.408	0.954
direct	0.198	0.408	0.954
agglo	0.684	0.187	0.723
graph	0.254	0.383	0.936
bagglo	0.234	0.392	0.939

Table 5: Evaluation of clustering results with SRSM and the new method for the second data set

α	SRSM			New method		
	Entropy	Mutual information	F measure	Entropy	Mutual information	F measure
0.30	0.205	0.405	0.953	0.133	0.438	0.971
0.35	0.141	0.434	0.970	0.125	0.442	0.974
0.40	0.155	0.428	0.965	0.122	0.443	0.975
0.45	0.175	0.418	0.960	0.137	0.436	0.971
0.50	0.179	0.417	0.959	0.150	0.430	0.967
0.55	0.172	0.420	0.962	0.186	0.414	0.956
0.60	0.182	0.416	0.956	0.161	0.425	0.963
0.65	0.196	0.408	0.952	0.174	0.419	0.959
0.70	0.202	0.406	0.951	0.188	0.413	0.955

Table 4 shows the evaluation of clustering results from CLUTO toolkit's algorithms. Table 5 shows the evaluation of clustering results of SRSM and the newly proposed method with different values of parameter α. The best evaluation in each quality measure values is shaded in Table 5.

For the WordNet semantic similarity based method, the evaluation of the clustering result was 0.363, 0.332, 0.894 for entropy, mutual information and *F* measure respectively.

The results show that clustering results of the newly proposed method are better than those of the other methods in all three evaluation measures.

6. Conclusions

The vector space model is widely used in the field of document clustering. It represents a document as a vector of terms. However, the simple VSM treats terms independent to each other and the semantic relationships between terms are not considered. Therefore, it reduces the effectiveness of document clustering methods. SRSM method and WordNet semantic similarity based method use the semantic relation between terms to improve the performance of document clustering. However, these methods have their own issues as we discussed in the previous sections. We proposed a new method that is a combination of SRSM and WordNet semantic similarity based method to solve these issues.

Our experiment results show that the quality of the clustering with the proposed method is better than the ones with SRSM and WordNet semantic similarity based method. Its clustering results are also better than results of other methods in the CLUTO toolkit.

In addition to WordNet, Wikipedia and Wiktionary are also promising tools for semantic relatedness measurement and analysis [22]. In our future work, we will exploit these tools to further improve document clustering methods.

References

[1] Y. Zhao and G. Karypis, "Hierarchical clustering algorithms for document datasets", Data Mining and Knowledge Discovery , 10 (2), pp. 141 - 168, 2005.

[2] I.S. Dhillon and D.S. Modha, "Concept decompositions for large sparse text data using clustering", Machine Learning, 42 (1-2), pp. 143-175, 2001.

[3] M. Steinbach, G. Karypis and V. Kumar, "A comparison of document clustering techniques", Proceedings of the KDD Workshop on Text Mining, 2000.

[4] Y. Li, S.M. Chung and J.D. Holt, "Text document clustering based on frequent word meaning sequences", Data and Knowledge Engineering, 64 (1), pp. 381-404, 2008.

[5] M. Mahdavi and H. Abolhassani, "Harmony K-means algorithm for document clustering", Data Mining and Knowledge Discovery, pp. 1-22, 2008.

[6] T.B. Ho and K. Funakoshi, "Information retrieval using rough sets", Journal of Japanese Society for Aritificial Intelligence, 13 (3), pp. 424-433, 1997.

[7] N.C. Thanh, K. Yamada and M. Unehara, "A Similarity Rough Set Model for document representation and document clustering", Journal of Advanced Computational Intelligence and Intelligent Informatics, 15 (2), pp. 125-133, 2011.

[8] W.K. Gad and M.S. Kamel, "Enhancing text clustering performance using semantic similarity", Lecture Notes in Business Information Processing, 24 LNBIP, pp. 325-335, 2009.

[9] L. Jing, M.K. Ng and J.Z. Huang, "Knowledge-based vector space model for text clustering", Knowledge and Information Systems, 25 (1), pp. 35-55, 2010.

[10] Z. Pawlak, "Rough sets", Int. J. of Information and Computer Sciences, 11 (5), pp. 341-356, 1982.

[11] R. Slowinski and D. Vanderpooten, "Similarity Relation as a basis for rough approximation", Advances in Machine Intelligence and Soft Computing, Vol.4, pp. 17-33, 1997.

[12] R. Slowinski and D. Vanderpooten, "A generalized definition of rough approximations based on similarity", IEEE Trans. on Knowledge and Data Engineering, 12 (2), pp. 331-336, 2000.

[13] T.B. Ho and N.B. Nguyen, "Nonhierarchical document clustering based on a tolerance rough set model", International Journal of Intelligent Systems, 17 (2), pp. 199-212, 2002.

[14] X.-J. Meng, Q.-C. Chen, and X.-L. Wang, "A tolerance rough set based semantic clustering method for web search results", Information Technology Journal, 8 (4), pp. 453-464, 2009.

[15] Princeton University, "About WordNet", WordNet, Princeton University. 2010, http://wordnet.princeton.edu.

[16] R. Rada, H. Mili, E. Bicknell and M. Blettner, "Development and application of a metric on semantic nets", IEEE Transactions on Systems, Man and Cybernetics, v (n), pp. 17-30, 1989.

[17] Z. Wu and M. Palmer, "Verbs semantics and lexical selection", Proceedings of the 32nd annual meeting on Association for Computational Linguistics (ACL '94), pp. 133-138, 1994.

[18] P. Resnik, "Using information content to evaluate semantic similarity", Proceedings of the 14th International Joint Conference on Artificial Intelligence, pp. 448-453, 1995.

[19] J. J. Jiang and D. W. Conrath. "Semantic similarity based on corpus statistics and lexical taxonomy", Proceedings of the 10th International Conference on Research in Computational Linguistics, Taipei, Taiwan, 1997.

[20] M. Lesk, "Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone", Proceedings of the 5th Annual International Conference on Systems Documentation, pp. 24-26, 1986.

[21] S. Banerjee and T. Pedersen, "An adapted lesk algorithm for word sense disambiguation using WordNet", CICLing ' 02: Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, pp. 136-145, 2002.

[22] T. Zesch and I. Gurevych, "Wisdom of crowds versus wisdom of linguists - Measuring the semantic relatedness of words", Natural Language Engineering, 16 (1), pp. 25-59, 2010.

[23] S. Patwardhan and T. Pedersen, "Using WordNet Based Context Vectors to Estimate the Semantic Relatedness of Concepts", Proceedings of the EACL 2006 Workshop Making Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together, pp. 1-8, Trento, Italy, 2006.

[24] G. Karypis, "CLUTO - A Clustering Toolkit", 2003, http://glaros.dtc.umn.edu/gkhome/cluto/cluto/download.

[25] ftp://ftp.cs.cornell.edu/pub/smart

[26] A. Strehl, J. Ghosh and R. Mooney, "Impact of similarity measures on web-page clustering", Proceedings of the 17th National Conference on Artificial Intelligence: Workshop of Artificial Intelligence for Web search (AAAI 2000), Austin, TX, pp. 58–64, July 2000.

Towards a Reference Model for Open Access and Knowledge Sharing, Lessons from Systems Research

Paola Di Maio
ISTCS.org
Edinburgh

Abstract

The Open Access Movement has been striving to grant universal unrestricted access to the knowledge and data outputs of publicly funded research. leveraging the real time, virtually cost free publishing opportunities offered by the internet and the web. However, evidence suggests that in the systems engineering domain open access policies are not widely adopted. This paper presents the rationale, methodology and results of an evidence based inquiry that investigates the dichotomy between policy and practice in Open Access (OA) of systems engineering research in the UK, explores entangled dimensions of the problem space from a socio-technical perspective, and issues a set of recommendations, including a reference model outline for knowledge sharing in systems research.

Keywords: **Systems Engineering Research, Knowledge Sharing, Reuse**

1. Introduction

The web provides without doubt the most efficient mechanism to exchange explicit knowledge, as long as this is codified and represented using appropriate formalisms and supporting artifacts. A wealth of research, platforms and technologies has been produced in recent decades much of it thanks to considerable public investment, yet despite the availability of good practices and no shortage of openly available knowledge sharing tools and platforms, much knowledge produced with taxpayer's money is still not shared, or only notionally shared, and there is no indication that the uptake of Open Access policies is actually monitored. The research aims to:

- identify policies and practices that regulate the explicit sharing of knowledge generated by publicly funded research in the UK, the body specifically in relation to systems engineering research,
- evaluate to what extent, and via which mechanisms and behaviors, the adoption of OA policies and knowledge sharing artifacts and processes are adopted, with specific focus for this study is systems engineering research in the UK
- devise examples of explicit knowledge models and artifacts to facilitate he codification and sharing of systems knowledge.

2. Contribution and Paper Organisation

This paper aims to identify and address a possible gap between the and the practice in Open Access in Scholarly research. It is organized as follows:

Knowledge sharing challenges: introductory discussion, and scope of the work.
Evidence and what works: overall research approach and Evidence Based Research, and an outline of the research plan.
Knowledge sharing behaviours and NECTISE: segmentation of the research field, and a case study, and ethnographic observations
Open access and knowledge sharing: filling the gap between two research strands
The surveying instruments: introducing Open Access Monitor and the Knowledge Audit Framework.
Preliminary findings: the initial results of this research to date.

3. Knowledge Sharing Challenges

Knowledge is one of the most valuable resources for individuals and organizations. Scholarly discussion on 'What is knowledge' (as opposed to information for example) are ongoing. For the purpose of this research the 'data-information-knowledge' classical distinction proposed in different versions by various authors is accepted (Ackoff; Bellenger; Sveiby; Davenport and Prusak). 'Knowledge sharing' is intended as making knowledge resources available on the web for free and unrestricted access, use and reuse. Despite decades of research and practice in knowledge management, knowledge sharing and reuse remains elusive, fragmented and compartmentalized (Mandl, et al). This in our hypothesis is due to systemic causes, which we address in our proposed approach. Several disciplines have been converging in recent years to facilitate and increase knowledge exchanges. Pervasive web based technologies have removed many of the physical barriers to knowledge sharing. However many challenges still inhibit optimal knowledge flows. This research targets the challenges associated with accessing knowledge that has been generated using public funding via public research councils in the UK: the UK is one of the countries perceived to be leading the ‘freedom of information good practice' and which has been spearheading ‘open access' since the early days, yet according to evidence gathered in our research, there are still many gaps in the practical uptake. In particular, since this research originated in the NECTISE research framework (Networked Enabled Capabilities for Systems Engineering) the current scope of the inquiry is primarily on systems engineering research in the UK, therefore constraining the focus of the analysis mostly to nationally funded research in Great Britain, however the research logic, as well as its instruments and methodology can be generalized and targeted to other domains and other countries, which we reserve to undertake in future work. In summary, the central problem this research tackles is that despite the existence of widespread open access policies which could appear *prima facie* to be in use, in the example of UK Research Councils, knowledge generated by Systems Engineering research using public funds is still not available to the public and sometimes not even to co-researchers.

4. Evidence of What Works

Knowledge reuse challenges can be examined under different disciplinary perspectives, but when tackled in combination, and considered ‘as a whole', systemic traits such as ‘entanglement’ emerge.

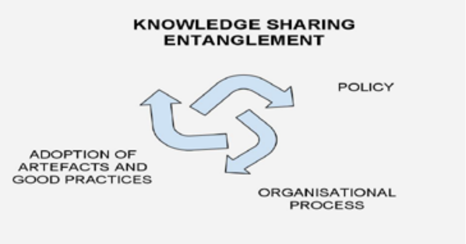

KNOWLEDGE SHARING ENTANGLEMENT
POLICY
ADOPTION OF ARTEFACTS AND GOOD PRACTICES
ORGANISATIONAL PROCESS

Image 1. Knowledge sharing entanglement

For example, knowledge sharing co-dependencies (entanglement) are addressed by relating two different dimensions of the problem space such as 'policy' and 'adoption of artifacts', constitutes the foundation of our mixed method research approach, as explained in related work (Di Maio, 2011) and illustrated schematically in image 1 above. The overall proposition that drives the inquiry is:

There is a gap between the existence of the adoption of open access policies 'in theory' (T) and 'in practice' (P)
from which the following questions and hypotheses are derived

Q1. How can the gap between T and P be identified?
H1. By gathering Evidence of the difference between T and P
Q2. How can the gap between T and P be measured?
H2. By devising indicators and parameters to evaluate the level of adoption
Q3. How can the gap between T and P be reduced?
H3. By devising and recommending appropriate measures and interventions

This paper provides an overview and a synthesis of findings obtained using different research methods, each contributing a piece of ’evidence' to help answer the question above, and to test the hypothesis. Evidence Based Research (EBR) emerges from a field known as 'Evidence based practice' (EBP):
Evidence-based practice (EBP) method in the behavioral and social sciences developed out of the evidence-based movement in medicine, which aims to inculcate in clinicians “the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients” (Sackett et al).

The rationale for EBR is rooted in clinical practice in the health and medical domains, however a methodology has grown out of it, that has been adopted by other social science disciplines. It is noted (Paynter) that :
While it may seem par for the course that professionals would use research to inform their practice, history is replete with examples of the opposite – practice based on the authority of their proponents rather than actual evidence of their efficacy. (Hatcher et al).

A typical EBP research process consists of the following steps:

(1) Formulate the question.
(2) Search for answers.
(3) Appraise the evidence.
(4) Apply the results.
(5) Assess the outcome

(Gray, 2004)

This research, described in more detail in the sections that follow, complies with the central tenets of what constitutes a ‘systemic review' method (EPPI) :
• Explicit and transparent methods are used following a standard set of stages.

• It is accountable, replicable and updatable.
• User involvement is built into the research design.

4.1 Research Design

The analytical part of study consists of two main research components, a Critical Appraisal of existing policies and legal instruments, and a systematic review of funded projects (the audits), that adhere to our inclusion criteria, described in more detail below.

1) A critical appraisal (evaluation) of Open Access and other Knowledge Sharing policies that guide or regulate academic practice in the UK (Davies), is aimed at answering the following questions

Exploratory evaluation: what policies are there? (Method: literature review, surveys and interviews with civil servants and experts)

Impact evaluation: do people know about these policies? (Method: audits, survey)

The steps undertaken in these research components are:

- Identify public research funding bodies

- Survey and analyze their policy and implementation strategies

2) Survey/Audits of the Field. To look for evidence showing that the policies are implemented as intended, or otherwise (test the hypotheses) a survey of a targeted sample of the population and projects of UK academic projects is undertaken. A survey instrument is designed to carry out data collection for this sample called OAM (www.openaccessmonitor.org), further specified in separate documentation linked on the site. The steps followed in this research component are:

- Select Cases to be audited (following the inclusion criteria)

- Gather data sets following OAM audit framework

- Analyse the findings according to multiple methods (qualitative, quantitative)

Inclusion criteria for the selection of cases in the current study are all the projects related to the target domain (in this study systems engineering research), UK-based and publicly funded through one or more UK research councils. The research concludes deriving from the findings a set of recommendations, which combine good practices with suggested interventions, such as policy integration and alignment, community involvement, and the adoption of suitable technical artifacts and knowledge models. It proposes a schematic reference model for shared knowledge representation, as well as other artefacts.

4.2 Motivation and Chain of Evidence

The initial motivation for the study was provided by NECTISE, a summary of the case is presented in the following section, as well as other observations, for which elements of ethnography were adopted. Academics (including principal investigators, researchers and postgraduate students) as well as practitioners showed little or no awareness of Open Access principles, which confirmed the findings of previous related studies (Swan). A series of interviews and email exchanges with experts followed to investigate various aspects of the problem space. The chain of evidence for this research is illustrated below:

NECTISE >>> Initial observations
Ethnography >>>Awareness of OA
Literature >>> Previous studies confirm observations
Survey, Interviews >>> Evidence from funding councils
Audits >>> Systemic survey of the field (ongoing)

4.3 NECTISE

The underlying, endemic problem tackled by this research is well illustrated by one of the exploratory cases that initially triggered, and largely still motivates, most of this research: a portion of the EPSRC funded NECTISE (www.nectise.com), a 4 million GBP research project awarded to a consortium of prestigious Universities for ‘networked capabilities in systems engineering', was allocated to investigate 'knowledge reuse'. As a doctoral training account holder (DTA) tasked with advancing the state of the art in ‘knowledge reuse and learning in networked capabilities research for systems engineering' and receiving doctoral research funding from the NECTISE funding pool, it was essential for this researcher to acquire and examine existing project knowledge before the state of the art could be advanced further. However, the only knowledge resources publicly available via the NECTISE website were a static list of published papers (not hyper linked, nor available via the site, just enumerated on an HTML page on the project website). Despite the fact that the project was publicly funded by EPSRC, due to contractual arrangements with industry partner BAE Systems. a private company which operates a policy of strict knowledge control, NECTISE never shared nor published any system diagrams, nor vocabularies or data dictionaries, and the research partners had to ask for permission to BAE before any decision could be taken. Although some of the papers linked on the project page could have been retrieved from scholarly repositories via web searches, they were mainly narratives and did not contain structured, systematic knowledge that could be re-used. An endless sequence of emails to obtain access to the knowledge artifacts related to the project between the doctoral researcher and entire research hierarchies of academics and officers in charge, generated no results - the 'target knowledge' was never obtained. No obvious open access project resources despite clear open access policies published by the funding body - prompted the question that motivates and justifies much of the current line of inquiry: if this research is publicly funded via EPSRC (Engineering and Physical Sciences Research Council), which like all other UK Research Councils embraces an 'open access' policy, why can't everyone, especially a researcher funded by the same project, access any of the project knowledge they need to carry our their professional or research task?

4.4 Knowledge Sharing Behaviours

An earlier pilot study, combined to international field work and an analysis of scholarly outputs, (Lee, Shiva) resulted in the identification of significant demographic differences that could contribute to shape the diversity of knowledge sharing behaviors For example a combination of factors, including Country, Job Role, Industry, Organisational Culture, can all impact to knowledge sharing attitudes and behaviors, a line of inquiry already partly explored in related research (Graf et al). While the theoretical part of this study is generalizable and domain independent (the research design and instruments can be modified to target different segments of the research field, or different research domains), given the constraint on resources it is necessary to narrow the current of scope of research to systems engineering research in the UK (see the gray cells in the table 1 below). The first research component, the critical appraisal of policy instruments, targets major research funding councils in the UK, considered in the context of related EU and international policies. The second part of the study, (the audits) has been targeted to systems engineering research projects in the UK

Table 1: Segmentation of the research field

	INDUSTRY	ROLE	SECTOR	POLICY
WORLD				
EU				
UK	systems engineering	researchers, research funding administrators	research	funding body, institution

4.5 Open Access for Knowledge Sharing

The regulations, legislation and policies that govern 'knowledge sharing' practices in academia and industry are an entangled web of instruments, characterized by the tension between a global cyber-culture on the one hand, that promotes knowledge sharing and the adoption of web based artifacts to facilitate free and unrestricted knowledge flows, and on the other hand strong commercial interests of a 'knowledge economy' that can exist only via restrictions to knowledge via intellectual property rights enforced through commercial contracts, which enable the materialisation of earnings from Knowledge Transfer activities, such as for example the sale of books, course fees, etc. For the purpose of the analysis, the regulatory landscape has been segmented as follows:

- International declarations (OECD, Budapest, Berlin)
- International directives (EU PSI 2003)
- National legislation (that apply in a single member state to all governing bodies, such as the FOI Act 2000)
- National policies of each governing body
- Knowledge Transfer policies

A more detailed exploration of each of the segments above is being reported in a separate paper (under review, as of August 2011), however for the purpose of this paper, a brief summary of each segment is provided below.

4.5.1 International Declarations

Open Access is the broad term that identifies a progressive movement and a series of initiatives that have gradually lowered the barriers to access publicly funded research outputs. There is a long and rich history that documents how this movement evolved thanks to the efforts of individuals, groups and collectives that has finally been embraced at least to some extent by institutions (Suber). Key initiatives include the Budapest Open Access Initiative, Berlin and Bethesda, which yielded slightly different definitions, however the classic definition of reference is:

"By "open access" we mean the free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles [...] (Budapest Open Access Initiative, 2002)"

Open Access to Scholarly Knowledge, has been a huge and growing movement, however it is noted that the initiatives above are not reflected in any legislation, and at the time of writing, no legislation exists that governs nor mandates the monitoring of open access publishing

4.5.2 International Directives

Public Sector Information has always been one of the main sources of primary data for many research activities and data centric services in modern economies, but thanks to the current explosion of web based technology applications and infrastructure that many more opportunities are opening up for a variety of stakeholders. The Council and the European Parliament adopted a Directive on the re-use of public sector information which deals with the way public sector bodies should enhance re-use of their information resources which, the EU says, is based on two key pillars of the internal market: transparency and fair competition (EU Council Directive of PSI Reuse 2003). The directive establishes minimum rules for the re-use of PSI throughout the European Union, but encourages Member States to go beyond minimum rules and to adopt open data policies, allowing a broad use of documents held by public sector bodies. Individual member states have adopted the directive with different legislative instruments and local variations (Implementation of the PSI directive). Interestingly, research institutions are excluded from the EU PSI directive with a comma in its Article 1

The Directive shall not apply to [...] *e) documents held by educational and research establishments, such as schools, universities, archives, libraries and research facilities including, where relevant, organisations established for the transfer of research results;*

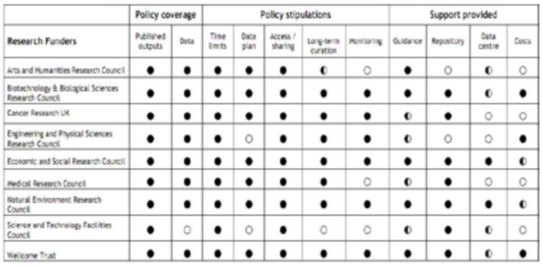
Neither the EU PSI Directive of 2003 nor the UK Re-use of Public Sector Information Regulations 2005 specifically define what public sector information is. However, both the EU Directive and the UK Regulations make clear it covers information and content that is held by public sector bodies that fall within the scope of the Directive Regulations and where the rights are held by the public sector body. (private email exchanges with the press office of the Office of The National Archive, July 2011). For the purpose of FOI legislation however (at least in the UK) universities are considered 'public authorities' and must comply with FOI legislation. There seems to be a contradiction between the definition of public authority of the FOI Act and of the Regulations derived from the EU Directive.

4.5.3 National Legislation (UK)

The most notable example of legislation aimed at making accessible and transparent public sector information, is the FOI Act. In the UK publicly funded Research Institutions and Universities are considered 'public authorities', and therefore PSI legislation applies (FOI Act 2000)In the UK, the FOI Act and the Regulation 1515, both aimed at increasing 'access to knowledge' seem to be conflicting in their definition of public authority.

4.5.4 National policies of individual governing bodies (UK)

At national level, each governing body responsible for a public sector, may adopt a different version of the relevant policy. For example in the UK, each of the five major research councils have a different position in relation to a)open access b) data sharing. In further work, we reserve to undertake a more detailed comparative analysis of the same. Image 2 below provides a notional comparison of the policies, based purely on what the policies documents state. A closer evaluation, supported by the findings of our audits, reported in a later section of this paper, reveals that some of policy coverage stated 'on paper' cannot easily be verified: for example, EPSRC states on paper that it monitors the policy implementation, while according both to what emerged in the NECTISE case and to other audits, EPSRC at the time our investigation started, did not keep a record of Open Access resources for each funded projects. (Research Log Entry[1]). Further work is currently being undertaken to obtain evidence from corresponding organisations of monitoring activities, however, the criteria and extent for monitoring are unclear.



	Policy coverage		Policy stipulations					Support provided			
Research Funders	Published outputs	Data	Time limits	Data plan	Access / sharing	Long-term curation	Monitoring	Guidance	Repository	Data centre	Costs
Arts and Humanities Research Council	●	●	●	●	●	◐	○	●	○	◐	○
Biotechnology & Biological Sciences Research Council	●	●	●	●	●	●	●	●	●	◐	●
Cancer Research UK	●	●	●	●	●	●	●	◐	●	○	○
Engineering and Physical Sciences Research Council	●	●	●	○	●	●	●	◐	○	○	●
Economic and Social Research Council	●	●	●	●	●	●	●	●	●	●	◐
Medical Research Council	●	●	●	●	●	●	○	◐	●	○	○
Natural Environment Research Council	●	●	●	●	●	●	●	●	●	●	◐
Science and Technology Facilities Council	●	○	●	○	●	○	○	◐	●	◐	○
Wellcome Trust	●	●	●	●	●	●	●	●	●	◐	●

Image 2: Curation Policies DCC Edinburgh[2]

4.5.5 Knowledge Transfer (KT)

Knowledge transfer (KT) can be used to describe the knowledge flows between different units, divisions, or organizations rather than individuals (Szulanski, Cappetta, Jensen), the emphasis of KT is on generating income from knowledge transfer activities, rather than maximising access to knowledge. KT is also defined as "the process through which one unit (e.g., group, department, or division) is affected by the experience of another" (Argote, Ingram). The EU Commission also states that it wants to move towards a position in which:

“knowledge transfer between universities and industry is

[1] http://fieldnote.posterous.com/knowledge-reuse-in-systems-engineering

[2] http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies

made a permanent political and operational priority for all public research funding bodies within a Member State, at both national and regional level". (Commission Recommendation)

As illustrated in Image 3, Knowledge Transfer principles consist of restricting access to knowledge, to allow for the commercial exploitation of knowledge resources, and generate income streams via the sale of educational materials (teaching), consulting services and licensing mechanisms - essentially in direct contrast with Open Access principles.

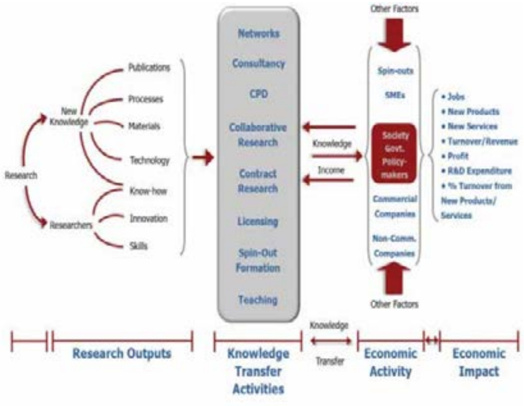


Image 3: Model of Knowledge Transfer within the Innovation Ecosystem (Source: University of Glasgow) In: Metrics for the Evaluation of Knowledge Transfer Activities at Universities (Unico Report)

Furthermore, the analysis of literature in the UK and EU reveals that Knowledge Transfer policies shape and mandate the knowledge exchange perceptions and behaviors at praxis level (Hauser). Intellectual Property clauses of commercial contracts part of 'Knowledge transfer programmes' restrict and constrain knowledge flows between academia and industry, effectively pre-empting Open Access policies to take hold. (Gardner, Fong, Huang). One of the asymmetries that become visible when contrasting of KT vs OA policies, is that the first are grounded in contract law, which is made firm in the law (contract law) while Open Access policies, at the time of writing, are still 'guidelines', so the first are prioritized due to their legal weight (Burnhill).The sections above provide an overview of the diverse set of intiatives, policies, and corresponding regulations that govern knowledge sharing practices, partly known as 'open access policies'. Findings and summary conclusions of this analysis of the landscape are presented in section 5 of this paper.

4.5.6 Auditing the Field

Despite the fragmentation of the regulatory landscape discussed above, each research funding council in the UK has their own 'open access policy' as reported in Image 2. The next step in our research process consisted of carrying out a systematic review of publicly funded projects, to see to what extent such policies were adhered to by principal investigators and their corresponding institutions. Although the digital curation community may not consider the distinction between information and knowledge, the so called 'knowledge level' (Newell) has different implications. An ad hoc Knowledge Auditing instrument was devised (KAF)[3] however this resulted to target 'too granular' level of knowledge - that is, KAF was devised using knowledge engineering principles aimed at specifying a high degree of formality of detailed technical knowledge. The KAF auditing process is illustrated in the image 4.

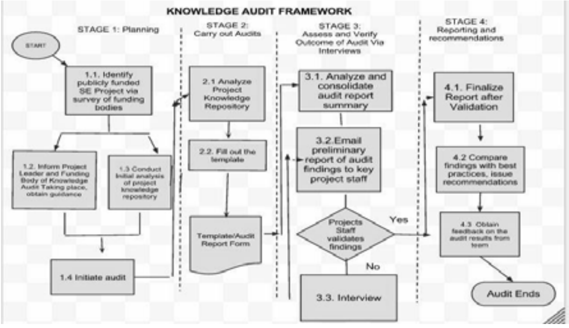

KNOWLEDGE AUDIT FRAMEWORK
STAGE 1: Planning
STAGE 2: Carry out Audits
STAGE 3: Assess and Verify Outcome of Audit Via Interviews
STAGE 4: Reporting and recommendations
START
1.1. Identify publicly funded SE Project via survey of funding bodies
1.2. Inform Project Leader and Funding Body of Knowledge Audit Taking place, obtain guidance
1.3 Conduct initial analysis of project knowledge repository
1.4 Initiate audit
2.1 Analyze Project Knowledge Repository
2.2. Fill out the template
Template/Audit Report Form
3.1. Analyze and consolidate audit report summary
3.2.Email preliminary report of audit findings to key project staff
Projects Staff validates findings
Yes
No
3.3. Interview
4.1. Finalize Report after Validation
4.2 Compare findings with best practices, issue recommendations
4.3 Obtain feedback on the audit results from team
Audit Ends

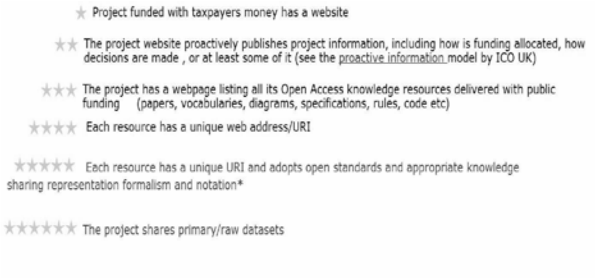
Image 4: Knowledge Audit Framework Process

After piloting KAF in the field, it emerged that the auditing instrument was over specified, for example it looked for systems specifications and diagrams, when it became clear that the majority of projects in the systems engineering research being audited did not even have a website, and of those which have a website, very few have links to accessible copies of deliverables and papers. Therefore a more generic, 'evidence based research' instrument evolved from KAF, called Open Access Monitor (OAM), a public version - slightly more polished instance of the data collection tool used in our audits for this research - is accessible on the web at http://www.openaccessmonitor.org. OAM consists of simple guiding principles, a process and data collection instruments (forms) and corresponding public data repositories to store the audited knowledge. It is designed to harvest a wider range of knowledge sharing standards, from the simplest form - 'does the project have a website'? - to more detailed, technical audit of knowledge sharing formalisms adopted - does the project open access resources have a unique web address (URI), and are they published using appropriate formalisms and

[3] KAF http://tinyurl.com/3oleaaf

notation?

The current version of OAM is a working prototype developed at ‘near zero cost', that is, using freely available development tools (Google apps). OAM evolved organically out of KAF, keeping the adopts its core process and inventorying mechanism, however, it uses an ‘abbreviated protocol' (a simpler and less granular inventory process). OAM uses different inventorying templates to gather evidence about existing Open Access Policies (Policy Monitor) and about how publicly funded projects embrace the policies (Project Monitor). OAM also encourages and supports public intervention by providing an open, publicly accessible record of civic interventions (i.e. it logs email requests sent to funding bodies when Open Access resources in relation to a given project are not found). OAM provides a unified environment to assist knowledge auditors 'score' publicly funded research projects according a simple star schema which constitutes a form of 'heuristic evaluation' (Nielsen, J; Porter et al). The star rating systems is modeled on the linked data star system (Berners Lee). OAM 's internal architecture (the process and the templates) and methodology are available as documentation, however the star system is illustrated in the image below.



★ Project funded with taxpayers money has a website

★★ The project website proactively publishes project information, including how is funding allocated, how decisions are made , or at least some of it (see the proactive information model by ICO UK)

★★★ The project has a webpage listing all its Open Access knowledge resources delivered with public funding (papers, vocabularies, diagrams, specifications, rules, code etc)

★★★★ Each resource has a unique web address/URI

★★★★★ Each resource has a unique URI and adopts open standards and appropriate knowledge sharing representation formalism and notation*

★★★★★★ The project shares primary/raw datasets

Image 5: Heuristic Assessment of KS via star system

Overall, OAM templates used in combination can help knowledge auditors answer the following questions:

1. What are the Open Access policies of each funding body? How is the implementation of these policies monitored?

2. Which Open Access knowledge resources are shared in the public domain for each publicly funded project? In addition, if a specific Open Access resource is not located, OAM encourages individual independent 'auditors' to write to the corresponding funding body, and to log such inquiry, related correspondence and responses in a public record.

4.5.7 The auditing sample

The target of the audits portion of the study are systems engineering research project in the UK, the funding council that specifically targets SE is the EPSRC, although other research councils such as the ATRC also funds large scale systems, they do not categorize 'systems engineering' as such. A comparative evaluation of different categorisation systems for different research councils points to the need to further harmonize, or at least map, the conceptual and categorization schemas for different councils, but we leave this discussion for a future work. Given the relatively contained number - approximately 100 - of systems engineering research projects funded by EPSRC that ended in 2010 and 2009 it was decided the sampling strategy was a 'census', ie, it did not require a selection of a subseet of the total sample, but given existing resources, and by recruiting volunteer auditors, they could all be audited. It should be noted that since the audits took place while OAM was still in development, only five of the six criteria were audited in our study (the sixth criterion was added later).

5. Results

Below a summary of preliminary findings to date, corresponding to each research components: policy evaluation (theory) evidence from the field (practice)

5.1 Policy Evaluation Findings

The policy assessment effort was initiated as part of this research with the goal of understanding what OA policies exist, and to what extent funding councils implement and monitor them. Different methods for policy evaluation were adopted in combination (Purdon et al). Outcomes of this evaluation point to the following conclusions:

1. The policy landscape is fragmented across different levels. For example, different policies addressed loosely different layers of the information management chain, for example: Data, Information, Knowledge.

2. There are different policies with different scopes and purposes, all targeting roughly the same ‘knowledge sharing' space, but which are not harmonized,

3. Some of the current legal provisions for the protection of Intellectual Property, and programmes such as 'Knowledge transfer' that restrict knowledge flows between academia and industry, could be in conflict with Open Access policies.

4. UK Funding bodies have Open Access policies in place, however they do not monitor, and when they do, they do not specify 'how' they monitor the implementation of OA policies.

5.2 Field Research Findings

The first observation, via the NECTISE case, provided the initial evidence that motivated the rest of the study. Pilot interviews were undertaken, which resulted in one of the important 'lessons learned': structured surveys and questionnaires may not result in honest, truthful replies. Respondents are intimidated by the technical jargon in use (what is a formalism? was one of the typical reactions) and were also reluctant to share openly their attitudes and behaviors. An ethnographic approach was therefore adopted from that point onward, and systems engineers were observed in the course of the three year in various occasions via

1. participation in UK and international workgroups such as INCOSE
2. participation in international systems engineering conferences and events (UK and international)
3. direct observation and participation in international systems engineering projects (Incose SEBOK)

The aim of ethnographic observation is to gain some understanding of:

- do systems engineering researchers know what is open access?
- do they know what the Budapest Initiative is?
- do they know what knowledge sharing policies govern their publicly funded research?

One of the ethnographic studies consisted of casual 'on the spot' interviews carried out in 2009, where academics (researchers and postgraduate students) were asked in their natural work environment, and in the context of routine 'reseach interest' type of conversations, whether they knew what is Open Access, and what is the Budapest Declaration; One of the studies was carried out on campus (an engineering faculty in the UK). Of 30 participants, selected randomly (were physically approached on campus when the opportunity arose) and anonymously (their names were not recorded) all answers were negative: nobody knew what Open Access is, nor what the Budapest Declaration is. The same ethnographic experiment was repeated across a variety or events, over a period of time, with slight variations in the results.

Table 2 summarises the type of events and dates, number of subjects who were approached and their responses to the three questions above.

Although of limited statistical significance, these result point clearly to lack of awareness of open access. The findings on our limited sample confirmed the outcomes of earlier reports (Swan).

Table 2: Summary of ethnographic study

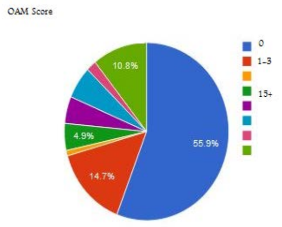
SETTING	DATE	Nr	Q1y	Q1n	Q2y	Q2n	Q3y	Q3n
walk in engineering campus (6 weeks, local)	2009	30	0	30	0	30	0	30
systems engineering networking meeting (1 day, national)	2010	24	0	24	0	24	0	24
Space symposium (local)	2010	14	0	14	0	14	0	14
Syseng Iternational conferences (4 days, international)	2010	30	4	26	0	30	5	25
Syseng National Conference (2 days, international)	2011	22	3	19	2	20	4	18

It was therefore decided that no further data was needed to demonstrate the 'lack of awareness' problem. Instead, a systematic survey of publicly funded projects in the SE domain was undertaken using OAM. Four initial pilot audits were carried out, which helped refine the monitoring instrument and fine tune the auditing procedures. A total of 100 EPSRC funded projects ended in 2009 and 2010 has been audited and 'scored' to date, with the following results: the majority of project audited did not have open access knowledge resources, or very few (<3), however the good news was that the third largest group of 11 audits scored very high (>14).

Table 3: OAM Scores

Number of Audits	Score
57	0
15	1
1	2
5	3
5	6
6	10
2	14
11	15

The pie chart below represents diagrammatically the figures in the table



Additional datasets with some variance in the inclusion criteria are being gathered to permit further analysis, however from the current findings, the following conclusions can be drawn:

1) The majority of projects audited did not have any or very few open access resources,

2) Almost all projects have some papers published that can be retrieved via web searches associated to the grant number

3) The third largest segment of the audited population (approx 10%) adheres to all good practices and knowledge sharing conventions (>14)

4) The quality, detail and sharing formalisms adopted by each project varies greatly and does not always depend on the existence of an explicit knowledge sharing policy. Web page, when they exist, are used mainly as promotion/marketing, and no consistent KS formalization is adopted.

5) Other factors, such as background of the grant holder and organisational culture may contribute to the level of granularity and knowledge sharing formalisms adopted.

6) the quality and type of knowledge shared in publicly funded systems engineering research tends to be high level, narratives (papers), however limited formalized and reusable system knowledge is routinely published and shared.

7) The minority of projects audited that adopt standard knowledge sharing practices (the notable exceptions) do so consistently and in compliance with good practices. Currently each of these projects is being used as a 'model of good practice', and studied more closely, to gather additional insights into outstanding KS behavior. In summary, the evidence gathered so far from field work points toward the following conclusions:

- Researchers in systems engineering are generally not aware of OA policies.

- Only a limited number of publicly funded projects complies with the policies of their funding bodies

- Open Access policies are underspecified and vague.

A number of other qualitative considerations that have emerged from the evaluation of the findings as a whole are currently being elaborated in a technical report that will be sent to all individuals and institutional representatives, and that will serve as the basis for further research.

6. Recommendations

Over the course of the study, initial evidence and findings were discussed and presented to various individuals in selected funding bodies and organizations, some of which are logged as research notes (research log, private correspondence). During the course of these exchanges, a press statement was issued by Research Councils UK (RCUK) and the Higher Education Funding Council for England (HEFCE)[4] announcing plans to work together to ensure greater open access to published research (Announcement, 25 May 2011). In particular, the EPSRC, the research council with which we have had intense correspondence exchanges for the last two years, circulated a policy update (Ryan, B) which states textually:

'EPSRC will monitor compliance with the policy as part of our normal assurance processes'

However there is no indication as to what the open access monitoring strategy going to be, and more importantly, no indication of what level of public resources are going to be devoted to this effort, in simple terms, it is not clear how much the monitoring of open access resources is going to cost the tax payer, nor how efficient and effective is going to be. One of the contributions of this research is a set of recommendations gathered partly from standard good practices, as we learn them from web science and knowledge engineering, and partly emerge from the empirical evaluation of the evidence gathered in the course of the study. These recommendations are grouped into four distinct categories:
- toward a reference model for knowledge sharing
- for governing councils, research funding bodies
- for institutions
- for individual researchers

'Open access' is a broad, boundary spanning complex socio-technical challenge, and the proposed recommendations are best adopted in combination: simple, cost effective measures articulated across the different levels of the problem space can yield systemic results.

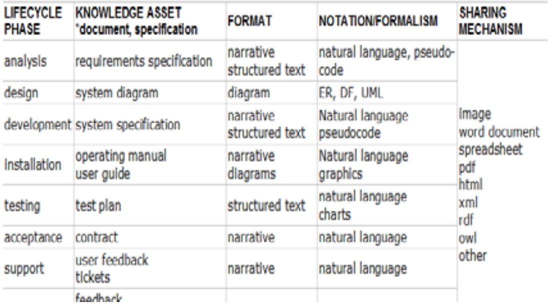
6.1 Towards a General Reference Model for Knowledge Sharing

To achieve optimal knowledge sharing potential of codified knowledge resources, such as technical knowledge, it is necessary to adopt appropriate conventions, formalisms and artifacts. Some of these conventions are well established, and have been encoded as a knowledge sharing star rating system for OAM . however no single knowledge schema exist that researchers can adopt when trying to make their outputs more useful, and more easily accessible. The rationale and workplan toward the development of a reference ontology and a shared vocabulary for the system engineering practice, is reported in a separate paper, (Di Maio, Proceedings of the ACM, 2011). The outcome of the knowledge and content analysis of knowledge

[4] http://www.rcuk.ac.uk/media/news/2011news/Pages/110525_1.aspx

resources in the systems engineering domain has resulted in a sample 'reference model' reported below, whereby the system development phases correspond knowledge artifacts, logically articulated, represented and shared using appropriate formalisms, notation and file formats. Similar domain dependent knowledge reference schemas can be developed and adopted in other fields of practice.

Reference Model of Knowledge Sharing in SE (Di Maio)



LIFECYCLE PHASE	KNOWLEDGE ASSET 'document, specification	FORMAT	NOTATION/FORMALISM	SHARING MECHANISM
analysis	requirements specification	narrative structured text	natural language, pseudo-code	image word document spreadsheet pdf html xml rdf owl other
design	system diagram	diagram	ER, DF, UML	
development	system specification	narrative structured text	Natural language pseudocode	
Installation	operating manual user guide	narrative diagrams	Natural language graphics	
testing	test plan	structured text	natural language charts	
acceptance	contract	narrative	natural language	
support	user feedback tickets	narrative	natural language	
	feedback			

The overall general recommendation - as well as one of the contributions of this research to the systems engineering research domain - is that basic traditional web knowledge sharing artifacts and core practices, such as the use of URIs for sharing knowledge resources, appropriately used address and resolve most knowledge sharing challenges. Furthermore, adopting a domain specific knowledge reference model, as illustrated in the table above, can mitigate at least in part the lack of more sophisticated shared codification standards.

6.2 Fact Checking

'Scientific knowledge' rests, above all, on facts, whereby science itself is about verifiability and reproducibility. This research is developed in the context of an engineering discipline, in particular systems, web and knowledge engineering, whereby engineering is intended as 'the practical application of science to commerce or industry' [Fox]. The ability to verify facts via gather evidence is essential to reason, make inferences, draw conclusions and essentially, to make informed decisions. On the web, which is the largest open, large scale distributed knowledge base, fact checking is particularly important to the accuracy of reasoning, which can be defined as the act or process of using one's reason to derive one statement or assertion (the conclusion) from a prior group of statements or assertions (the premises) by means of a given method [Clarke]. The validity of 'knowledge' requires it to be verified or verifiable, with some exceptions that may be satisfied with theoretical assumptions. Fact checking is adopted routinely in investigations (research) in providing evidence (legal/making the case) and in decision making (to reduce over reliance on assumptions). It is recommended that when sharing knowledge on the web, the mechanism to provide verifiable evidence is to use hyper links to corresponding documents, which can be either HTML or RDF. In related work, the linked data model is explored as a possible formalization for fact checking [5]

6.3 For policy makers and funding bodies

The fragmented state of heterogeneous policies and legislation can be confusing, and even lead to contradictory practices, as identified in the relevant section of this paper. Although it is acceptable to have multiple policies, it would be advisable a certain level of cohesion, integration and alignment between them.

a) An open access policy management strategy should enable dual track, i.e. encourage compliance from the bottom up (self archiving) but also encourage funding bodies and regulators to implement the policy via regulatory measures (mandates) and above all monitor compliance with the policy

b) bridge the current fragmentation between data, information and knowledge policies, and establish a firm 'correspondence' between the policy and the mandates on the one hand, which can be called the social and organisational aspects of knowledge sharing, and the adoption of the knowledge sharing artefacts, conventions and standards, that can be defined as the technical aspects, because the two are facets of the 'same coin', as shown diagrammatically in the illustration below.

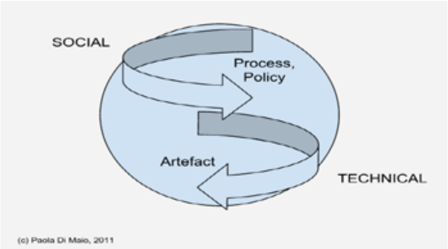


Image 6: Socio Technical Approach to Open Access

c) devise and implement an overall integrated Open Access policy monitoring strategy which should be in line, and where possible extend, the guidelines provided by international directives, such as EU PSI Directive 2003.

d) Leverage the community: it is expected that budgetary considerations will play a role in how effectively the monitoring of Open Access policies implementations will be. If carried out manually, and without use of ICT, the

[5] Provenance and Linked Data Workshop, SICSA, University of Edinburgh 2011

cost of monitoring policy implementation could exceed its benefits. However, if a simple automated policy monitoring process is implemented by mandate, say via an open web service such as OAM, the burden of monitoring could be distributed across the research community or even crowd-sourced which would reduce the material costs of a much needed monitoring to almost zero.

d) issue clear guidelines as to what level of data, information and knowledge should be made freely accessible to by mandate, and which levels can be protected by patents and copyright to allow research outputs to benefit from commercialization opportunities and economic gain via Knowledge Transfer agreements

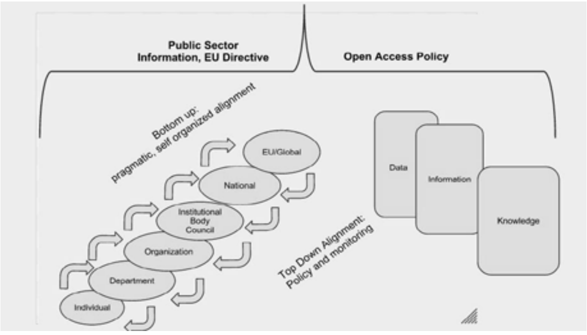


Image 7: Intervention: Integration and alignment of the fragmented policies regulating the space

d) consider legislation. At the moment, the provisions for commercial knowledge transfer are entered contractually, whereby contracts are legally binding instruments entered enforced by contract law. Open access policies are still operated as guidelines only, and carry no legal, binding weight. The relation between open access policies and knowledge transfer agreements is strongly asymmetrical in the law, and favors the latter.

6.4 For research institutions

Institutions, as large bureaucratic organisations, tend to be 'passive', and to follow directions issued from 'the top' by governing bodies. When a policy carrying strategic implications for the advancement of science at global and national level, such as the policy for open access to scholarly publications, it is necessary for everyone in the research supply chain to wholeheartedly embrace it. What good is a policy emitted by a funding body, if no institution adheres to it?Institutions have primary responsibilities toward the public at large, as well as toward the public funding councils, and the research community. Their responsibility is to understand the open access framework, and to pass it on to their entourage. The primary recommendations for research institutions are as follows:

a) embrace the culture of knowledge sharing. this often implies a disruptive overhaul of pre-constituted knowledge hierarchies, and it cannot be achieved overnight

b) provide regular training about knowledge sharing and where necessary technical support for researchers

c) issue guidelines and recommendations as to what optimal knowledge sharing practices are, including recommending the adoption of existing artifacts and good practices, and stimulate the innovative development of new ones.

6.5 For Researchers

In contemporary networked society, self governance, as well as the active participation of individuals in all governance practices of institutions, is encouraged, but this cannot happen without the researchers understanding the political and practical implications of information policies.

1. Publish often, as often as possible, and do not wait for results to be complete and exhaustive, share your findings early, update the findings with progress reports.

2. Share your knowledge and data using standard good practices, contribute to the development of the same.

3. Favor, where possible, working for institutions transparent and compliant with good practices and support open access

4. Contribute to the active evangelization and monitoring of open access in your research environment, and become a point of reference for your community.

7. Contribution, and Future Work

This research so far claims the following contributions:
- the first systematic review of open access in the systems engineering research
- the first 'evidence based research' contributed to systems engineering research
- contributes the novel concept and example of 'heuristics evaluation' to knowledge sharing research (the star system)

Additional data cross validation for ancillary quantitative analysis of the findings is currently being undertaken. Future work includes a wider study using OAM in other domains and countries, a contribution to public consultations both in the UK and the EU.

8. Conclusion

This paper presents the rationale, methodology and some of the findings and recommendations of a study aimed at filling the gap between Open Access theory and practice. It introduces OAM, a near zero cost public environment to support the monitoring of open access policies and presents an example of 'reference model' for knowledge sharing in systems engineering.

9. Acknowledgments

This research is partly funded with EPSRC Grant nr. EP/D505461/1 for the project. “Network Enabled Capability Through. Innovative Systems Engineering (NECTISE). Gratitude to many people who are providing essential feedback and guidance in handling the complex intricacies and dependencies of the diverse components that make this research possible

Paola Di Maio holds a BA Hons (1994) and an MSc (2000), is a research analyst for Cutter.com, independent expert for the European Research Agency (REA). She works as a standards evaluator and research advisor, is a Research Associate at Institute of Socio-technical Complex Systems in the UK, lectures internationally.

References

Ackoff, R. L. "From Data to Wisdom", Journal of Applied Systems Analysis, Uniwersytet Warszawski, 1989

Argote, Ingram, "Knowledge transfer: A Basis for Competitive Advantage in Firms". Organizational Behavior and Human Decision Processes, 2000

Bellenger, G. "Creating Knowledge Objects", 2004, accessed online August 2011
http://www.reusability.org/ read/chapters/ wiley.doc.

Berners Lee, Tim - Web data star system, 2010
http://lab.linkeddata.deri.ie/2010/star-scheme-by-example/

Burnhill, P. Director of EDINA, Private conversation Edinburgh August 2011

Calise, M., de Rosa, R. and Fernandez X. "Electronic Publishing, Knowledge Sharing and Open Access: a new environment for political science", European Political Science, Palgrave Journal 2010

Davenport Prusak, "Working knowledge: how organizations manage what they know", Harvard Business School Press 1998

Davies, P. "Policy Evaluation in the United Kingdom", 2004, accessed online August 2011
http://www.nationalschool.gov.uk/policyhub/docs/policy_evaluation_uk.pdf

Di Maio, P. "Toward a semantic vocabulary for systems engineering" WIMS '11 Proceedings of the International Conference on Web Intelligence, Mining and Semantics. ACM New York, NY, USA 2011

EPPI http://eppi.ioe.ac.uk/cms/Default.aspx?tabid=89,

Gartner, Fong, Huang, "Measuring the impact of knowledge transfer From Public Research Organisations", Int. J. Learning and Intellectual Capital , 2010
http://www.mcgill.ca/files/senate/Impact_of_Knowledge_Transfer.pdf

Graff, M. G., Davies, J. & McNorton, M. "Cognitive style and cross-cultural differences in internet use and computer attitudes" European Journal of Open and Distance Learning, 2004

Gray J.A. Muir "Evidence-based Healthcare" 2nd Edition, Churchill Livingstone, Elsevier, Philadelphia, 2004
Hatcher, S., Butler, R. and Oakley-Browne, M., "Evidence-based Mental Health Care", Elsevier Churchill Livingstone, Edinburgh 2004

Hauser, H "The Current and Future Role of Technology and Innovation". Centres in the UK, Crown Copyright, 2010

Jones, S. HATII Glasgow, Personal Correspondence
Research log entry, 2011

Lee, Shiva "A Novel Approach to Knowledge Sharing in Software Systems Engineering" Fourth IEEE International Conference on Global Software Engineering, 2009

Mandl, Gruber and Renkl Ch 8. "Misconceptions and Knowledge Compartmentalization", Advances in Psychology, Elsevier, 1993

Newell, "The Knowledge Level", Presidential Address, American Association for Artificial Intelligence. Stanford University, 1980

Nielsen, J. "Heuristic evaluation". In Nielsen, J., and Mack, R.L. Usability Inspection Methods, John Wiley & Sons, 1994

Robin A. Paynter, "Evidence-based research in the applied social sciences", Reference Services Review, 2009
www.emeraldinsight.com/0090-7324.htm

Porter et al "Concept learning and heuristic classification in weak-theory domains"- Artificial Intelligence - Elsevier.1990

Purdon, Lessof, Woodfield and Bryson, "Research methods for policy evaluation", Department for Work and Pensions, Crown Copyright 2001

Suber, P. Timeline of the Open Access Movement, 2009
http://www.earlham.edu/~peters/fos/timeline.htm

Sveiby, K.E. "The New Organizational Wealth:Managing and Measuring Knowledge-Based Assets" Berrett-Koehler Publishers, 1997

Swan. A, "Key Concerns within the Scholarly Communications Process", Key Perspectives 2008

Unico Report
http://ec.europa.eu/invest-in-research/pdf/download_en/library_house_2008_unico.pdf

Wang, Noe "Knowledge sharing: A review and directions for future research" Human Resource Management Review, Elsevier, 2010

Artificial Neural Network for Transfer Function Placental Development: DCT and DWT Approach

Mohammad Ayache[1], Mohamad Khalil[2] and Francois Tranquart[3]

**[1] Department of Biomedical, Faculty of Engineering, Islamic University of Lebanon
Khalde Highway, BP 30014, Lebanon**

**[2] Faculty of Engineering, Lebanese University
Tripoli, Lebanon**

**[3] Faculty of Medicine, University of Tours
Tours, France**

Abstract

The aim of our study is to propose an approach for transfer function placental development using ultrasound images. This approach is based to the selection of tissues, feature extraction by discrete cosine transform DCT, discrete wavelet transform DWT and classification of different grades of placenta by artificial neural network and especially the multi layer perceptron MLP. The proposed approach is tested for ultrasound images of placenta, resulting in 75% success rate of classification using DCT and 92% using DWT. The method based on multi resolution decomposition analysis and on supervised neural network technique MLP, seems a good method to study the transfer function of placental development in ultrasound.

Keywords: *Placenta, discrete cosine transform, discrete wavelet transform, neural network, MLP.*

1. Introduction

An ultrasound diagnostic system has become an important and popular diagnostic tool due to its wide range of applications. The non-invasive, non-destructive nature of ultrasound, real-time imaging and portable system are the advantages over the medical imaging system such as X-ray, CT and MRI. Due to these advantages of ultrasound diagnostic system, it has been extensively used in the medical profession.

An ultrasound diagnostic system, which has been widely used in the field of medical diagnosis, transmits ultrasound signals to a target and receives echo signals reflected from the target to form an ultrasound image. The ultrasound diagnostic system generally uses a wide bandwidth transducer to transmit and receive ultrasound signals. The received signals are beam-formed and passed in the several processes to form an ultrasound image. Diagnostic ultrasound is generally perceived by users and patients as a safe technique with no adverse effects. Since ultrasound is so widely used in pregnancy, it is essential for all practitioners to ensure that its use remains safe.

However it is difficult to differentiate between normal and abnormal tissues on the basis of ultrasound images from the placenta. The placenta is a vascular organ formed in the uterus during pregnancy, consisting of both maternal and embryonic tissues and providing oxygen and nutrients for the fetus and transfer of waste products from the fetal to the maternal blood circulation. A study of placental development in ultrasound is needed to evaluate the different grades of normal placental maturation and assess the transfer function and detect abnormalities, particularly those that may be responsible for a premature birth or intrauterine growth retardation. Grannum [1], due to a visual approach, classifies the placental grades into 4 different grades. The goal of the research work in the present paper is to classify automatically the grade of the placenta based on advanced image processing techniques such that discrete wavelet transform DWT and discrete wavelet transform DCT. In the literature, there are different methods that can be used to extract the images in order to make the diagnosis. Wavelet transforms [2] Fourier transform [3], discrete cosine transform [4], [5] and continuous wavelet transform [6] can be used to extract parameters from ultrasound images. Statistical parameters are used for the texture analysis [7]. Wavelet is used for the detection of the micro calcification mammograms [8]. Discrete wavelet transform DWT and discrete cosine transform DCT are used in our study for image processing and feature extraction. Artificial neural network ANN and especially the multi layer perceptron MLP is used for the image classification. ANN is widely used for medical image classification [9] [10].

2. Material, Methods and Problem Statement

One hundred normal pregnant volunteers were scanned in Bretonneau Hospital (Tours, France). The gestational age varied from 21 weeks to 36 weeks. Conventional ultrasound units (Sequoia 512, Siemens; Voluson 530, GE) were used to obtain two dimensional ultrasound images. Only anterior and lateral placentas were considered to avoid any interference with the baby and to allow a more valuable characterization. Data were stored during a sweep of the transducer on the maternal abdomen. The ultrasound images were acquired while applying the probe in an orthogonal orientation to the chorionic plate. All images were acquired in a gray scale mode. Non linear imaging technique was used to acquire the images.

The placental scans were then examined and graded according to the Grannum classification and to the classification developed from our experience in ultrasonic study of the placenta. Four relatively distinct phases of placental maturation have been indentified based on changes which occur in three separate zones: chorionic plate, placental substance and basal layer (figure 1).

Grade 0: the chorionic plate is smooth. The substance of the placenta is homogeneous and devoid of echogenic densities. The basal layer area is also devoid of echogenic densities.

Grade 1: the chorionic plate will appear to have subtle undulations but it may sometimes be difficult to appreciate this if the fetus is closely approximated to the plate. The substance of the placenta may well contain echogenic densities that are randomly dispersed in the substance of the placenta. These are linear in shape, with their long axis parallel to the long axis of the placenta. The basal layer of the placenta is still devoid of echogenic densities.

Grade 2: As the placenta matures, the echogenic densities become more numerous in number and more dense. The chorionic plate may appear more markedly indented with linear echogenic densities perpendicular to the chorionic plate which extend into the substance of the placenta but not all the way to the basal layer area.

Grade 3: this configuration describes the appearance of the placenta being divided into compartments, the cotyledons. The chorionic plate again is indented.

We have to not that a placenta may have It should be noted that a given placenta may have simultaneously more than one grade if different sections are examined. In evaluating each scan in this series, the assigned grade corresponded to the most mature portion of the placenta assessed. It is obviously important to visualize as much placental tissue as possible.

In normal pregnancies, it was reported that grade 1 appears at 31 weeks of gestation, grade 2 at 36 weeks and grade 3 at 38 weeks. But there is no relationship between the gestational age and the grade classification. [11], [12], [13], [1].

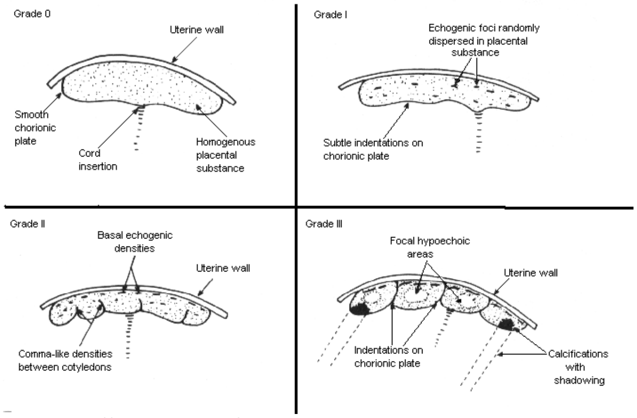

Grade 0
Uterine wall
Smooth chorionic plate
Cord insertion
Homogenous placental substance
Grade I
Echogenic foci randomly dispersed in placental substance
Subtle indentations on chorionic plate
Grade II
Basal echogenic densities
Uterine wall
Comma-like densities between cotyledons
Grade III
Focal hypoechoic areas
Uterine wall
Indentations on chorionic plate
Calcifications with shadowing

Figure 1: different grades of placenta

The aim of our studies is to establish an objective placental classification directly from the images. For that, for each image we extract the characteristic parameters and then we apply a classification method to determine the grade. This is a preliminary study to determine the transfer function placental development using advanced image processing techniques and artificial neural network.

3. Discrete Wavelet Transform

The wavelet transform is a very useful tool in the analysis of images. The theory and methods of wavelet analysis are widely presented in books [14], [15]. In this paper, discrete wavelet analysis is used instead of the continuous wavelet analysis. The discrete wavelet analysis is based on the concept of multiresolution analysis (MRA) introduced by Mallat [2]. This characterization of the wavelet transform allows the study of an image from the coarse resolution to the fine resolution and the extraction of information in any levels of decomposition. The (MRA) can be implemented with a two channel filter bank using quadrature mirror filters. The algorithm applies a one- dimensional high and low pass filtering step to both the rows and columns to the input image. Each filtering step is followed by subsampling which results in change in scale. Transforms in image processing are two-dimensional, so we need a few comments on how we implement a separable transform. When a two dimensional transform is separable, we can calculate it by applying the corresponding one dimensional transform to the column first, and then to the rows (Figure 2). At each decomposition level there are four different output images. An approximation of the input image and three detail images. The information contained in the output subbands of the DWT are:

- LL coefficients which correspond to a low pass filter to rows, followed by low pass filter to columns.

- HL coefficients which correspond to a low pass filter to rows, followed by high pass filter to columns.
- LH coefficients which correspond to a high pass filter to rows, followed by low pass filter to columns.
- HH coefficients which correspond to a high pass filter to rows, followed by high pass filter to columns. When a separable transform is applied, only the LL coefficients may need further decomposition. When this decomposition is done at many levels, we get the subband decomposition in Figure 2.

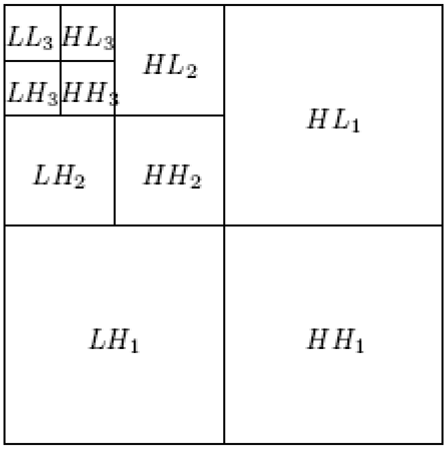


Figure 2: passband structure for a two dimensional subband transform with three levels

4. Discrete Cosine Transform

Like other transforms, the Discrete Cosine Transform (DCT) attempts to decorrelate the image data. After decorrelation each transform coefficient can be encoded independently without losing compression efficiency. This section describes the DCT and some of its important properties [21][22][18].

The objective of this section is to study the efficacy of DCT on placental images. This necessitates the extension of ideas presented in the last section to a two-dimensional space. The 2-D DCT is a direct extension of the 1-D case and is given by

$$C(u,v) = \alpha(u)\alpha(v)\sum_{x=0}^{N-1}\sum_{y=0}^{N-1} f(x,y) cos\left[\frac{\pi(2x+1)u}{2N}\right] cos\left[\frac{\pi(2y+1)v}{2N}\right] (1)$$

for u,v =0,1,2,…,N −1 and $\alpha(u)$ and $\alpha(v)$ are

$$\alpha(u) = \left\{\begin{matrix} \sqrt{\frac{1}{N}} for\ u = 0 \\ \sqrt{\frac{1}{2N}} for\ u \neq 0 \end{matrix}\right\}(2)$$

$$\alpha(v) = \left\{\begin{matrix} \sqrt{\frac{1}{N}} for\ v = 0 \\ \sqrt{\frac{1}{2N}} for\ v \neq 0 \end{matrix}\right\}(3)$$

The 2-D basis functions can be generated by multiplying the horizontally oriented 1-D basis functions with vertically oriented set of the same functions [18][19][20]. The basis functions for N = 8 are shown in. Again, it can be noted that the basis functions exhibit a progressive increase in frequency both in the vertical and horizontal direction. The top left basis function of results from multiplication of the DC component in Figure 3 with its transpose. Hence, this function assumes a constant value and is referred to as the DC coefficient.

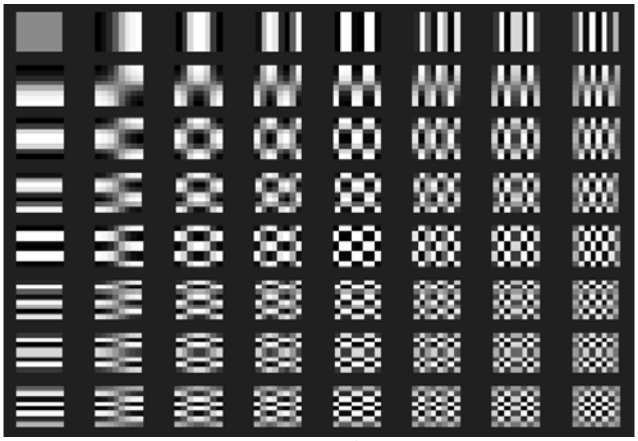


Figure 3: Two dimensional DCT basis functions (N = 8). Neutral gray represents zero, white represents positive amplitudes, and black represents negative amplitude[18].

5. Artificial Neural Network

After we extract the parameters from the DWT and DCT, an artificial neural network is used for the purpose of classification.

After the calculation of parameters, an artificial neural network is used to classify the images into different grades. A neural network is a general mathematical computing paradigm that models the operations of biological neural systems. The multilayer perceptron (MLP) is by far the most well known and most popular neural network among all the existing neural network paradigms. This type of neural network is known as supervised network because it requires a desired output in order to be learned [16], [17]. The goal of this type of network is to create a model that correctly maps the input to the output using historical data so that the model can then be used to produce the output when the desired output is unknown.

This work chooses MLP and its optimization to train and test the data of image registration. The recognition performance of the MLP network will highly depend on the structure of the network and training algorithm. In the current study, back projection algorithm selected to train the network. The number of nodes in input, hidden and output layers will determine the network structure. Furthermore, hidden and output nodes have activation function that will also influence the network performance. The best network structure is normally problem dependent, hence structure analysis has to be carried out to identify the optimum structure. Figure 4 shows a complete general schematic for the MLP network used in our study.

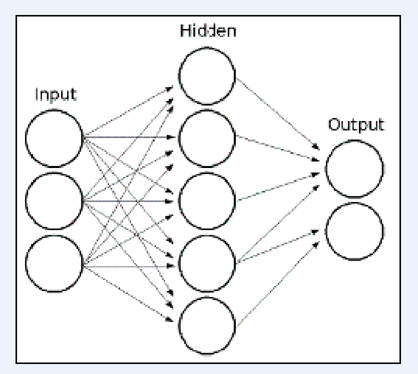


Figure 4: General schematic for the MLP

6. Results

6.1 Image Modeling Using DCT

Figure 5 shows the ultrasound images for different grades of placenta obtained in our experiment. After we select the region of interest, a DCT is applied to the selected region. Normally, in image compression, the DCT is applied to blocks 8*8 or 16*16. Because our purpose is not to compress an image, but an image classification after extracting the parameters and the position of the placenta is not always in the same place, the DCT was applied to the entire selected region. Figure 6 shows the application of the DCT to the different grades of placenta.

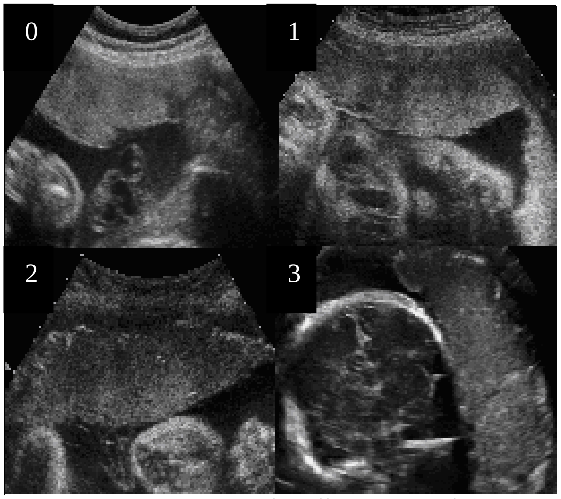


Figure 5: ultrasound images for different grades of placenta

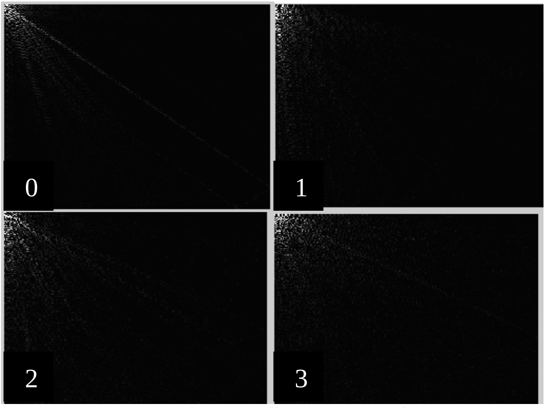


Figure 6: DCT application for the different grades of placenta

The figure 6 shows that the majority of energy is in the components of low frequencies with negligible values at high frequencies. If the image contains more details or heterogeneity, as the placenta, more the high frequency coefficients are important. For example, the grade III which normally is the grade the more heterogeneous contains more coefficients important at high frequencies. To quantify this difference of frequency in the different grades of placenta, statistical parameters are extracted after the application of the DCT. The extracted parameters are the mean, the standard deviation, the variance and the kurtosis.

6.2 Classification Using MLP

The neural network used for the purpose of classification is composed of one input layer contained 4 neurons correspond to the extracted parameters, one hidden layer with 5 neurons, and one output layer with 4 outputs correspond to the 4 different grade of placenta.

The activation functions used in our neural network are the following:

- Input layer: without activation function (no calculation at the level of this layer).
- Hidden layer: Sigmoid function

$$\left(\varphi(x) = \frac{1}{1+e^{-x}}\right) \quad (4)$$

- Output layer: Sigmoid function

$$\left(\varphi_o(x) = \varphi(x)\right) \quad (5)$$

Or Hyperbolic function

$$\left(\varphi_o(x) = \tanh\left(x/2\right) = \frac{1-e^{-x}}{1+e^{-x}}\right) \quad (6)$$

40 images /100 are used to train the neural network and the others are used for testing the classification to study the performance of the used neural network. A confusion matrix (table 1) shows the result of the classified image and the misclassified images.

Table 1: confusion matrix for the DCT classification

Grade	0	I	II	III	Miss classified
0	**8**	**4**	**2**	0	**6**
I	0	**11**	**4**	0	**4**
II	0	**0**	**12**	**3**	**3**
III	0	0	**2**	**13**	**2**

Table 1 shows that 15 images /60 are miss classified in the different grades of placenta. We note that 75% of images are classified correctly. The percentage of error is 25%.

6.3: Image Modeling using DWT

After the selection of the region of interest, a discrete wavelet transform at level 3 is applied to the ultrasound images. This technique allows to decompose each region of interest in different regions, which correspond to: an approximation region at the last level (level 3), and others details regions at different levels. Figure 7 shows the wavelet decomposition based on the multi resolution analysis at level 3.

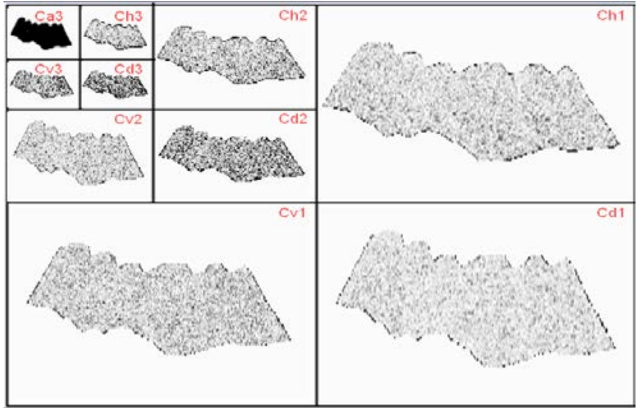


Figure 7: wavelet decomposition of placental images at level 3.

After the application of the discrete wavelet transform, the variance is extracted from each region. A vector of parameters is ready for the neural network input.

6.4: Classification using MLP

The neural network used is the same of the model applied in the DCT transform technique. There is only one difference in the number of neurons on the input layer which equal to 10 inputs correspond to the 10 parameters extracted from different regions after the application of the DWT.The number of images used for training and for classification in DWT is the same used in the DCT technique (40 images for training and 60 images for classification). The confusion matrix below (table 2) shows the result of classification obtained after the modeling of the images by DWT and the classification using the MLP.

Table 2: confusion matrix for the DWT classification

Grade	0	I	II	III	Miss classified
0	**12**	**3**	**0**	0	**3**
I	0	**14**	**1**	0	**1**
II	0	**0**	**14**	**1**	**1**
III	0	0	**0**	**15**	**0**

Table 2 shows that 5 images /60 are miss classified in the different grades of placenta. We note that 92% of images are classified correctly. The percentage of error is 8%.
The choice of wavelet depends on the type of analysis and the image taken into consideration. Experimentally, different types of wavelets are tested. The Daubechies were the best. The table below (table 3) shows the classification obtained for different type of wavelets using MLP.

Table 3: Different types of Waveletes tested in MLP

Type of Wavelet	Classification Rate
Haar	85%
Symlet	80%
Coiflet	85%
Daubechies	92%

7. Discussions and Conclusions

Two techniques were used to extract the detailed parameters to guide us toward a better classification. Decomposition in Discrete cosine transform has first been applied to the placental images. Generally, this technique used to decompose image in different frequencies and applied to blocks 8*8 or 16*16. As the placental location is not the same in all images, it was impossible to apply this technique to blocks. For this reason, it was applied to the total images then statistical parameters were extracted from the images. The results obtained with MLP neural network give a 75 % of correct classification. A small improvement was achieved by this technique.
Now, to extract parameters more effective or relevant for the classification, a more detailed decomposition analysis is requested. The wavelet transform is a technique used for frequency decomposition while preserving the temporal location. The discrete wavelet transform is base on multi resolution analysis at different level. Experimentally, level 3 was chosen and applied to the images. Extraction of variance was performed for each region of decomposition. Since this method is based on multi resolution analysis, that is meaning, we use more the details of the images; this method is effective to highlight the classification of different grades of placental maturation. The results obtained by the MLP showed a correct classification of 92%.

The technique based on multi resolution decomposition analysis and on supervised neural network technique MLP, seems a good method to study the transfer function of placental development in ultrasound. In fact, the choice of number of grades is correlated to the Grannum classification. We can obtain intermediate grades between grade 0 and 1 or between 1 and 2, but it will not help us in terms of diagnosis. Obtaining a grade in a given week of gestation may give an important significant. For this reason, an automatic classification of grades should help to characterize better the placental risk of pregnancy.

References

[1] Grannum, P.A., Hobbins, J.C.: The placenta. Radiologic Clinics of North America Vol. 20, No. 2, June (1982) 353-365.

[2] Mallat, S. (1989). A theory for multiresolution signal decomposition: The wavelet representation. IEEE Transactions on PAMI, 11, 674–693.

[3] Feleppa, E. J., Kalisz, A., Melgar, J. B. S., Lizzi, F. L., Liu, T., Rosado, A.L., et al. (1996). Typing of prostate tissue by ultrasonic spectrum analysis. IEEE Transactions on Ultrasonics Ferroelectrics and Frequency Control, 43(4), 609–619.

[4] Kurnaz, M.N., Dokur, Z., & O¨ lmez, T. (2001). Segmentation of ultrasound images by using an incremental self organized map. In: 23rd Annual international conference of the IEEE-EMBS, Vol. 3 (pp. 2638–2640).

[5] Mehmet Nadir Kurnaz, Z¨ umray Dokur, Tamer ¨ Olmez. An incremental neural network for tissue segmentation in ultrasound images. Computer methods and programs in biomedicine 8 5 (2007) 187–195.

[6] Kurnaz, M.N., & O¨ lmez, T. (2007). Tissue Segmentation in ultrasound images by using genetic algorithms. (Article in press).

[7] Haralick, R. M., Shanmugan, K., & Dinstein, I. (1973). Texture feature for image classification. IEEE Transactions on Systems Man and Cybernetics, 3, 610–621.

[8] Zhang, W., Yoshida, H., Nishikawa, R. M., & Doi, K. (1998). Optimally weighted wavelet transform based on supervised training for detection of microcalcifications in digital mammograms. Medical Physics, 25(6), 949–956.Erickson J., http://www.ics.uci.edu/%7Eeppstein/gina/voronoi.html, 1996.

[9] Serhatlioglu, S., Hardalac, F., Guler I. Classification of transcranial Doppler signals using artificial neural network. J Med Syst.2003 Aprl; 27(2):205-14.

[10] L. F. A. Campos, A. C. Silva, A. K. Barros: Independent Component Analysis and Neural Networks Applied for Classification of Malignant, Benign and Normal Tissue in Digital Mammography. Methods of Information in Medicine 2007 46 2: 212-215.Drysdale S., *Voronoi Diagrams Applications from Archeology to Zoology*, Regional Geometry Institute, 1993.

[11] Grannum, P.A., Berkowitz, R.L.,andHobbins, J.C.: The ultrasonic changes in the maturing placenta and their relation to fetal pulmonic maturity. Am. J. Obstet. Gynecol., 133:915-922, 1979.

[12] Grannum, P.A.: The placenta. Clinics in diagnostic Ultrasound, Vol.25, 203-219, 1989.

[13] Grannum, P.A.: Ultrasound examination of the placenta. Clinics in obstetrics and Gynecology, Vol.10, No.3, 459-73, 1983.

[14] Chui C.K., An introduction to wavelet, academic Press, 1992.Chui C.K., An introduction to wavelet, academic Press, 1992.

[15] Teolis A., Computational signal processing with wavelets, Brikhäuser, Boston 1998.

[16] Neural Networks: A comprehensive Foundation (2nd Edition). Simon Haykin.

[17] D.E. Rumelhart et al, 1986. Learning representations by back- propagating errors, Nature, 1986.

[18] W. B. Pennebaker and J. L. Mitchell, "JPEG – Still Image Data Compression tandard,"Newyork: International Thomsan Publishing, 1993.

[19] G. Strang, "The Discrete Cosine Transform," *SIAM Review*, Volume 41, Number 1, pp. 135-147, 1999.

[20] R. J. Clark, "Transform Coding of Images," New York: Academic Press, 1985.

[21] A. K. Jain, "Fundamentals of Digital Image Processing," New Jersey: Prentice Hall Inc., 1989.

[22] A. C. Hung and TH-Y Meng, "A Comparison of fast DCT algorithms," *Multimedia Systems*, No. 5 Vol. 2, Dec 1994.

Mohammad Ayache obtained a bachelor of engineering in biomedical from the Islamic University of Lebanon. He received the DEA in Signals and Images in biology and medicine from the University of Angers, France in 2004. He received the Ph.D degree in medical Image Processing from the University of Tours, France, in 2007. He is the coordinator of the department of biomedical at the faculty of engineering at the Islamic University of Lebanon. His research interests include advanced neural networks software development and advanced signal and image processing techniques.

Mohamad Khalil obtained an engineering degree in electrical and electricity from the Lebanese University, faculty of engineering, Tripoli, Lebanon in 1995. He received the DEA in biomedical engineering from the University of Technology of Compiegne (UTC) in France in 1996. He received his Ph.D from the University of Technology of Troyes and University of Technology of Compiegne in France in 1999. He received his HDR (Habilitation à diriger des recherches) from UTC in 2006. He is currently Professor at the Lebanese university, Faculty of engineering, section 1 Tripoli. He is researcher in the Lebanese University and his current interests are the signal and image processing problems: detection, classification, analysis, representation and modeling of non stationary signals, with application to biomedical signals and images. He is Responsible of the research axis (Image and Data Transmission) in LaMA Laboratory, Faculty of sciences, Lebanese university. Director of the center Azm of research in biotechnology and applications in the Lebanese university.

François Tranquart (MD, PhD) was born in Chatillon sur Indre (France) in 1958. He is General Manager of Bracco Suisse SA, Geneva Research Center and Manufacturing site since 2010. This site is dedicated to the development of new ultrasound contrast agents as well as developing solutions to support adequate *in vitro* and *in vivo* use of those agents. As a full professor of biophysics, he worked previously at Tours University Hospital as head of Ultrasound Medical Dept (with 100 exams performed daily), head of Centre For Innovation (new developments) and head of a research team (INSERM U930) dedicated to the development of new ultrasound approaches including microbubbles. His main interest is for new diagnostic agents or applications as well as new therapeutic developments involving the use of bubbles. He has presented more than 400 papers in national and international conferences, written more than 120 papers and participated to 20 books.

Time Series based Temperature Prediction using Back Propagation with Genetic Algorithm Technique

Shaminder Singh[1], Pankaj Bhambri[2] and Jasmeen Gill [3]

**[1] CSE, Punjab Technical University, BCET
Ludhiana, Punjab 141206, India**

**[2] CSE, Punjab Technical University, GNE
Ludhiana, Punjab 141002, India**

**[3] CSE, Punjab Technical University, RIMT-IET
Mandi Gobindgarh, Punjab 147301, India**

Abstract

Temperature prediction is a temporal and time series based process. Accurate forecasting is important in today's world as agricultural and industrial sectors are largely dependent on the temperature. Due to non-linearity in climatic physics, neural networks are suitable to predict these meteorological processes. Back propagation integrated with genetic algorithm is the most important algorithm to train neural networks. In this paper, in order to show the dependence of temperature on a particular data series, a time series based temperature prediction model using integrated back propagation with genetic algorithm technique is proposed. In the proposed technique, the effect of under training and over training the system is also shown. The test results of the technique are enlisted along with.

Keywords: *Artificial Neural Networks, Back Propagation Algorithm, Genetic Algorithms, Time Series Prediction.*

1. Introduction

Forecasting is a phenomenon of knowing what may happen to a system in the next coming time periods [4]. Temporal forecasting, or time series prediction, takes an existing series of data $x_{t-n}, \ldots, x_{t-2}, x_{t-1}, x_t$ and forecasts the data values $x_{t+1}, x_{t+2}, \ldots, x_{t+m}$. The goal is to observe or model the existing data series to enable future unknown data values to be forecasted accurately [11]. As weather is a continuous, data-intensive and dynamic process, the parameters required to predict temperature are enormously complex such that there is uncertainty in prediction even for a short period [8]. These properties make temperature forecasting a formidable challenge. The property of artificial neural networks that they not only analyze the time series data but also learn from it for future predictions makes them suitable for time series based temperature forecasting. Neural networks provide a methodology for solving many types of non-linear, time based problems that are difficult to be solved through traditional techniques [11]. Hence these characteristics of neural networks guided this research work to use them for the prediction of the meteorological processes.

Inspired by the brain, neural networks are an interconnected network of processing elements called neurons. Neural networks learn by example i.e. they can be trained with known examples. One of the most popular training algorithms in the domain of neural networks used so far, for temperature forecasting is the back propagation algorithm (BPN). As the algorithm suffers from many problems, attempts have been made by various researchers to solve these problems using genetic algorithms [2], [5]. Due to the temporal nature of weather processes, time series prediction has been introduced, but, no advancement beyond feed-forward neural networks trained with integrated back propagation and genetic algorithm has revolutionized the field. Therefore, much work still waits.

So, the motivation of this work is firstly, to develop time series based temperature forecasting model using hybrid neural network approach i.e. integrated BPN with GA and secondly, to estimate the effects of under-training and over-training the model.

The remainder of the article is organized as follows. Section 2 introduces the artificial neural networks. Next, a brief description of the time series predictor, neural network predictor and data series partitioning is given in Section 3. The details of the time series based temperature forecasting based on integrated BP/GA technique are shown in Section 4, followed by results in Section 5. Finally, conclusions are summarized in Section 6.

2. Artificial Neural Networks

Artificial Neural Network can be defined as a pool of simple processing units (neurons) which communicate among themselves by means of sending analog signals. These signals travel through weighted connections between neurons. Each of these neurons accumulates the inputs it receives, producing an output according to an internal activation function. This output can serve as an input for other neurons, or can be a part of the network output [3].

Back propagation is a systematic method of training multilayer artificial neural networks. It is built on sound mathematical base. The back propagation is a gradient descent method in which gradient of the error is calculated with respect to the weights for a given input by propagating the error backwards. The combination of weights which minimizes the error function is considered to be a solution of the problem [1].

Although Back propagation algorithm is an efficient technique applied to classification problems, system modeling, adaptive robotics control, but it suffers from local minima problem, scaling problem, long training time etc [1].

Genetic Algorithms developed in 1970 by John Holland, are computerized search and optimization algorithm that mimic the principle of natural genetics and natural selection. Genetic Algorithms perform directed random searches through a given set of alternatives to find the best alternative with respect to given criteria of fitness [6].

To eliminate the problems of back propagation algorithm, integrated BP/GA technique was developed. This technique is an efficient approach if only the requirement of a global searching is considered. It is good at global search (not in one direction) and it works with a population of points instead of a single point. Also it blends the merits of both deterministic gradient based algorithm BP and stochastic optimizing algorithm GA [1].

Neural Networks have been widely used as time series forecasters: most often these are feed-forward networks which employ a sliding window over the input sequence. Typical examples of this approach are market predictions, meteorological and network traffic forecasting [10]. Two important issues must be addressed in such systems: the frequency with which data should be sampled, and the number of data points which should be used in the input representation.

3. Time Series Prediction

A time series is a sequence of vectors, $\boldsymbol{x}$(t), t = 0,1,… , where t represents elapsed time. For simplicity we will consider here only sequences of scalars, although the techniques considered generalize readily to vector series. Theoretically, *x* may be a value which varies continuously with t, such as a temperature. In practice, for any given physical system, *x* will be sampled to give a series of discrete data points, equally spaced in time.

Work in neural networks has concentrated on forecasting future developments of the time series from values of x up to the current time. Formally this can be stated as: find a function f: $R^N \rightarrow R$, such as to obtain an estimate of x at time t + d, from the N time steps back from time t, so that:

$$x\,(t+d)=f\,(x(t),\,x(t-1),\mathrm{K},\,x(t-N+1)) \qquad (1)$$

and

$$x(t+d)=f(\mathbf{y}(t)) \qquad (2)$$

where $\mathbf{y}(t)$ is the N - ary vector of lagged *x* values

Normally d will be one, so that *f* will be forecasting the next value of *x* [10].

Time series forecasting has several important applications. One application is preventing undesirable events by forecasting the event, identifying the circumstances preceding the event, and taking corrective action so the event can be avoided. Another application is forecasting undesirable, yet unavoidable, events to preemptively lessen their impact. At this time, the sun's cycle of storms, called solar maximum, is of concern because the storms cause technological disruptions on Earth and other meteorological processes like interplanetary shocks, volcano eruption, cyclones, earth quakes, tsunamis etc. can also be predicted with the help of time series based forecasting. Also the daily weather forecasting of any place useful for agricultural purposes can be predicted through it. Finally, many people, primarily in the financial markets, would like to profit from time series forecasting. Whether this is viable is most likely a never-to-be-resolved question. Nevertheless many products are available for financial forecasting [11].

The standard neural network method of performing time series prediction is to induce the function f using any feed-forward function approximating neural network architecture using a set of N-tuples as inputs and a single output as the target value of the network. This method is often called the *sliding window technique* as the N-tuple input slides over the full training set. Figure 1 gives the basic architecture [10].

One typical method for training a network is to first partition the data series into three disjoint sets.

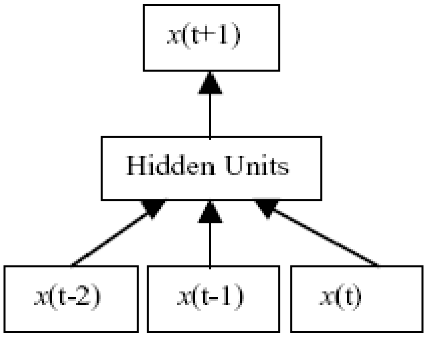


Fig. 1 The standard method of performing time series prediction using a sliding window of, in this case, three time steps.

These sets are: the training set, the validation set, and the test set. The network is trained (e.g., with back propagation) directly on the training set, its generalization ability is monitored on the validation set, and its ability to forecast is measured on the test set. A network's generalization ability indirectly measures how well the network can deal with unforeseen inputs, in other words, inputs on which it was not trained [11]. A network that produces high forecasting error on unforeseen inputs, but low error on training inputs, is said to have over-fit the training data. Over-fitting occurs when the network is blindly trained to a minimum in the total squared error based on the training set. A network that has over-fit the training data is said to have poor generalization ability.

To control over-fitting, the following procedure to train the network is often used:

1. After every 'n' epoch, sum the total error for all examples from the training set.
2. Also, sum the total squared error for all examples from the validation set. This error is the validation error.
3. Stop training when the trend in the error from step 1 is downward and the trend in the validation error from step 2 is upward.

When consistently the error in step 2 increases, while the error in step 1 decreases, this indicates the network has over-learned or over-fitted the data and training should stop. When using real-world data that is observed (instead of artificially generated), it may be difficult or impossible to partition the data series into three disjoint sets. The reason for this is the data series may be limited in length and/or may exhibit non-stationary behavior for data too far in the past (i.e., data previous to a certain point are not representative of the current trend) [9], [11].

4. Time Series based Temperature Prediction

In this section, the features of temporal data to the integrated BP/GA technique are introduced. The data used in this research are the daily weather data for the Ludhiana city of Punjab (India). The data in the un-normalized form have been collected from the "Meteorological Department of Punjab Agriculture University, Ludhiana (Punjab)" of the year 2009. The first thirty days data (month of January, 2009) have been used in this research. For training, the first twenty five days data have been used and next five days data have been used for testing purposes [2].

Before feeding the data to the network, it is to be converted to normalized form so as to provide improved performance or otherwise the use of original data to network may cause convergence problem. All the weather data sets were, therefore, transformed into values between 0 and 1 through dividing the difference of actual and minimum values by the difference of maximum and minimum values [7].

The neural network architecture along with the inputs required to feed to the network to perform time-series based temperature prediction and the outputs are shown below in fig. 2.

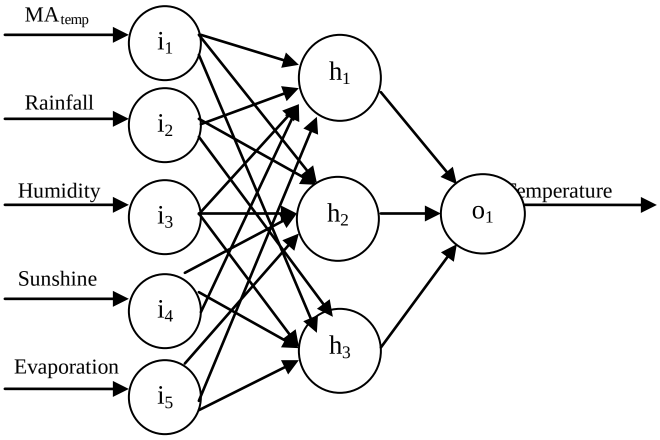


Fig. 2 Neural Network Architecture for the Proposed Model.

In this research, 5-3-1 neural network architecture has been used for BP/GA technique. The number of input neurons is 5 representing the moving average of mean air temperature of the previous temperature data, daily rainfall, relative humidity, sunshine and evaporation for the day for which the mean air temperature is to be predicted, the number of hidden neurons is 3 for processing and the number of outputs is 1 representing the weather variable i.e. mean air temperature to be forecasted.

The proposed time series based model starts with the collection of weather related data, selecting the weather parameters to be forecasted, extracting the relation between the different weather parameters, formation of training data set (containing inputs and outputs) and test data set (containing inputs).

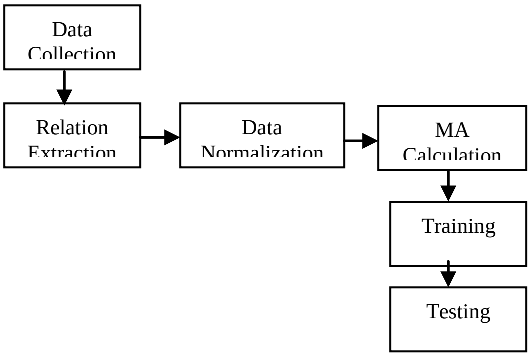


Fig. 3 Methodology for the Proposed Model.

For performing the time series prediction, a sliding window of size 5 has been moved over the full data set to obtain the moving average. This along with the dependent parameters act as an input to the system and have been used to train the network.

For training the network, an initial population of chromosomes is randomly generated. Then the weights are extracted from each chromosome depending upon the number of genes a chromosome is having. The cumulative error and the fitness are calculated over the inputs obtained from forecasting data. Then the crossover operator is applied for preparing the new population. This process is repeated till the stopping condition has been reached [1], [2].

5. Results and Discussions

The temporal weather forecasting technique based on BP/GA technique has been implemented by taking different population sizes.

Table 1: Selection of appropriate NN architecture

Population Size	*Hidden Neurons*	*Iterations*	*MAPE*
30	1	110	1.10
60	2	140	0.86
90	3	220	0.42
120	4	282	1.47
150	5	425	1.85

For each value of population, the program has been executed and the error has been calculated. Table 1 shows the variations in population size, number of neurons in hidden layer and the corresponding mean absolute percentage error values.

From the table 1, it is clear that the MAPE value is the lowest corresponding to population size 90 and number of hidden neurons as 3. So the present setup will use this population size for further research. The error vs. iteration graph corresponding to population size 90 is shown in fig. 4.

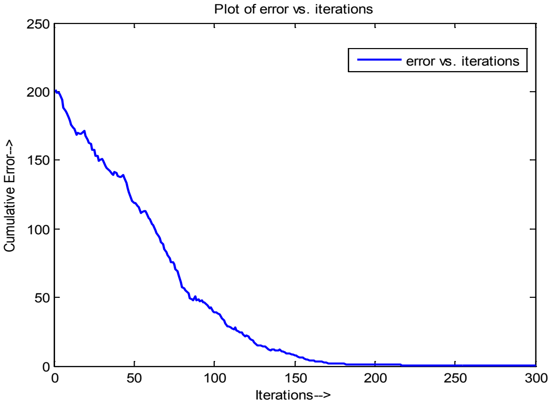


Fig. 4 The cumulative error v**a**lues corresponding to iterations for population size 90.

The error values corresponding to mean air temperature are shown in fig. 5 along with the Series 1, Series 2 and Series 3 of the inputs are shown. Error values are shown after 200 epochs, 400 epochs and 600 epochs. Clearly Series 3 shows the minimum error values in all the cases and it shows the lowest value after 400 epochs.

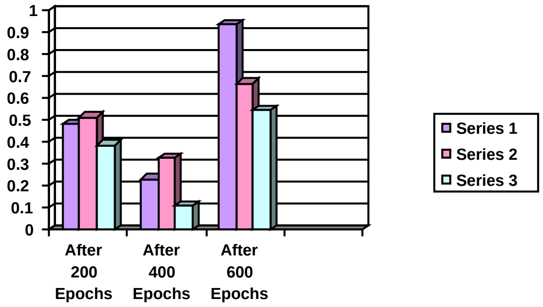


Fig. 5 Temporal prediction for mean air temperature.

Below is shown the actual prediction of mean air temperature using the proposed method along with the desired output as recorded with the help of instruments in fig. 6.

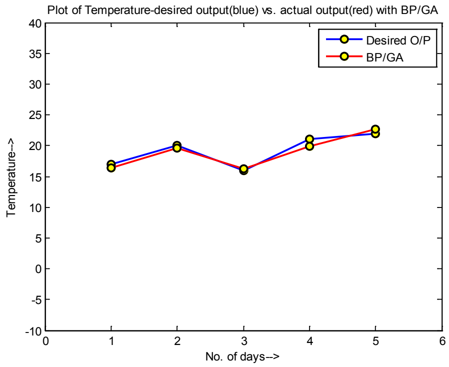


Fig. 6 The five-day mean air temperature prediction using the proposed model.

It is clearly shown in the fig. 6 that the time series based temperature prediction model using integrated BP/GA technique is suitable to predict the temperature. Secondly, dependence of weather parameters on the time series data is clearly shown in fig. 5. For the same weather parameter, the error values come out to be different when the network is trained with different data series. Thirdly the effect of under training the network through 200 epochs and over training the network through 600 epochs is easily visible as the error values are lowest after 400 epochs.

6. Conclusions

From the analysis above, it is easy to observe the compensability between time series based BP/GA technique and the back propagation alone. The proposed technique can learn efficiently by combining the strengths of GA with BP. It is good at time series data, global search (not in one direction) and it works with a population of points instead of a single point. Also it blends the merits of both deterministic gradient based algorithm BP and stochastic optimizing algorithm GA. Hence the use of the time series based temperature prediction model using integrated BP/GA technique is proposed.

References

[1] S. Rajasekaran and P. Vijayalakshmi, Neural networks, Fuzzy Logic and Genetic Algorithms, New Delhi: Prentice Hall of India, 2004.

[2] J. Gill, B. Singh and S. Singh, “Training Back Propagation Neural Networks with Genetic Algorithm for Weather Forecasting”, 8th IEEE International Symposium on Intelligent Systems Subotica Serbia, 2010, vol. 8, pp. 465-469.

[3] A. Azadeh and et al., “Integration of ANN and GA to Predict Electrical Energy consumption”, 32nd IEEE Conference on Industrial Electron., IECON, Paris, France, 2006.

[4] P. Sarangi and et al., “Short Term Load Forecasting using Artificial Neural Network: A Comparison with Genetic Algorithm Implementation”, J. of ARPN Engineering and Applied Sciences, vol. 9, 2009.

[5] D. E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Addison - Wesley. Reading, MA, 1989.

[6] H.S. Rao, V.G. Ghorpade and A. Mukherjee, “A genetic algorithm based back propagation network for simulation of stress-strain response of ceramic-matrix-composites”, J. of Computers and Structures, vol. 84, no. 5-6, 2006.

[7] I. Maqsood, M.R. Khan and A. Abraham, “Weather Forecasting Models Using Ensembles of Neural Networks”, 3rd International Conference on Intelligent Systems, Design and Applications Germany, 2003, pp. 33-42.

[8] A. Abraham et. al., “Soft Computing Models for Weather Forecasting” J. of Applied Sciences and Computations, USA, vol. 11, no. 3, pp. 106-117, 2004.

[9] M. Peter and Roth, “Temporal Pattern Recognition with Neural Networks” Pattern Recognition & Machine Learning, 1998.

[10] R. J. Frank, N. Davey and S. P. Hunt, “Time Series Prediction and Neural Networks”, J. of Intelligent and Robotic Systems, vol. 31, no. 1-3, 2001.

[11] E. A. Plummer, “Time Series Forecasting with Feed-forward Neural Networks: Guidelines and Limitations”, M.S. thesis, Department of Computer Science, The Graduate School of The University of Wyoming, Laramie, USA, 2000.

Colored Satellites Image Enhancement Using Wavelet and Threshold Decomposition

Dr. Muna F. Al-Samaraie[1] and Dr. Nedhal Abdul Majied Al Saiyd[2]

1. Management Information System Department
Faculty of Economics and Administrative Sciences
Al-Zaytoonah University, Amman, Jordan

2. Computer Science Department
Faculty of Information Technology,
Applied Science University, Amman-Jordan

Abstract

For decades, several image enhancement techniques have been proposed. Although most techniques require profuse amount of advance and critical steps, the result for the perceive image are not as satisfied. In this study, we proposed a new method to enhance the satellite image which using intelligent aspect of filtering and describe multi-threshold technique with an additional step in order to obtain the perceived image. In this way, several simple filters can be combined to form a more efficient and more flexible context dependent filter. This paper proposes a new idea for edge enhancement using smoothing techniques. Various smoothing technologies are explored and implemented for comparing their performance with various threshold values.

Comparison is done by qualitative and quantitative approaches. From the results, specific filtering is only applied to the region for which it is suitable. We also evaluate the image quality. The proposed method shows dramatically increase in pixel distribution throughout the range of RGB. The result of this research is also beneficial in terms of geographical views due to the process which determined the difference appeared on each area.

Keywords: Image enhancement, sharpening, wavelet, threshold decomposition and satellite images.

1. Introduction

In the modern information system, digital images have been widely used in a growing number of applications. The effort on edge enhancement has been focused mostly on improving the visual perception of images that are unclear because of blur. In general, the popular edge enhancement filtering is carried out with the help of traditional filters [1, 2 and 3]. But these filters do have some problems, especially while enhancing a noisy image.

Noise removal and preservation of useful us information are important aspects of image enhancement. A wide variety of methods have been proposed to solve the edge preserving and noise removal problem. Recently, researchers have focused their attention on nonlinear smoothing techniques in the spatial domain. Most of these techniques are local smoothing filters, which replace the center pixel of the neighborhood by an average of selected neighbor pixels.

Mainly focusing on the clarity of the image and the number of computations done for enhancing the image, we developed a novel approach. The edge enhancement done by smoothing filters decreases the complexity and also increases the quality of the image [5]. The basic aim of edge enhancement is to modify the appearance of an image to make it visually more attractive or to improve the visibility of certain features specially the satellite images. The edge enhancement technique enhances all high spatial frequency detail in an image, including edges, lines and points of high gradients. In this approach, the details of edges in an image can be obtained by subtracting a smoothed image from the original [4]. This subtractive smoothing method has been used as the simplest way to obtain high spatial frequency image and this method of edge enhancement makes the image brighter and real edges are detected.

In spite of all these efforts, none of the proposed operators are fully satisfactory in real world applications. They do not lead to satisfactory results when used as a means of identifying locations at which to apply image sharpening. In this paper, the enhancement is applied through a framework of threshold decomposition. This has two advantages: it reduces the edge detection to a simple binary process; and it makes the estimation of edge direction straightforward. Edge detection and direction estimation may be carried out by identifying simple

patterns, which are closely related to the Prewitt operators [6].
The previous methods stated can work under certain circumstances; meanwhile, it is in accurate to use the method in case of low contrast image or low resolution image [12]. Another method was proposed lately on satellite image enhancement which proposed an additional step by enhancing the brightness of the image before working on edge detection. However, the process shows no statistical results in the research.
Therefore, the quality evaluation was still un-identical in order to compare the results. In this study we proposed a novel approach satellite image sharpening, we developed new algorithms in intensity evaluation and compare its quality with its original version. The processes were composed of image brightness, edge detection and the standard deviation of the image intensity performed by the Peak Signal to Noise Ratio (PSNR).
The structure of the paper is arranged as follows: section 1 included the introduction and section 2 included the methodology of the proposed scheme. The proposed method is explained with many details in Section 3. Section 4 included the results. Conclusions are shown in Section 5

2. Methodology

2.1 Image enhancement

Image enhancement is a process principally focuses on processing an image in such a way that the processed image is more suitable than the original one for the specific application. The word “specific” has significance. It gives a clue that the results of such an operation are highly application dependent. In other words, an image enhancement technique that works well for X-ray topographic images may not work well for satellite images.
The technique falls in two categories on the basis of the domain they are applied on. These are the *frequency* and *spatial* domains. The frequency domain methods works with the Fourier Transforms of the image. The term spatial domain refers to the whole of pixels of which an image is composed of. Spatial domain methods are procedures that operate directly on the pixels. The process can be expressed in Eq. (1):

$$g(x, y) = T[\, f(x, y)] \qquad (1)$$

Where *f(x, y)* is the input image, *g(x, y)* is the processed image, and *T* is an operator on *f* defined over some neighborhood of *(x, y)* [7]. A number of enhancement techniques exist in the spatial domain. Among these are histogram processing, enhancement using arithmetic, and logical operations and filters.
For the study of each and every filter, we have considered the following algorithms for implementation:

Mean Filter: Mean filtering is simply the process of replacing each pixel value in an image with the mean (average) value of its neighbors, including itself [4,11]. This is simply done using 3*3 kernel.

Median Filter: The median is calculated by first sorting all the pixel values from the surrounding neighborhood in numerical order and then replacing the pixel being considered with the middle pixel value [11]. This is also implemented using 3*3 kernels.

Mode Filter: Mode filtering simply involves the replacing of each pixel value in an image by the mode value of its neighbors, including itself [11]. This is also implemented by 3*3 kernels.

Circular Filter: Circular filter is implemented using the product of original matrix and convolution mask provided [11]. A 5*5 kernel is used here.

Pyramidal Filter: Pyramidal filter is implemented using the product of the original matrix and convolution mask provided [11]. A 5*5 kernel is used here.

Cone Filter: Cone filter is implemented using the product of original matrix and convolution mask provided [11]. A 5*5 kernel is used here.

2.2 Threshold Decomposition

Threshold decomposition is a powerful theoretical tool, which is used in nonlinear image analysis. Many filter techniques have been shown to ‘commute with thresholding’. This means that the image may be decomposed into a series of binary levels, each of which may be processed separately. These binary levels can then be recombined to produce the final grayscale image with identical pixel values to those produced by grayscale processing. Hence a grayscale operation may be replaced by a series of equivalent binary operations.
The first threshold decomposition framework for image processing was introduced by [8]. This was capable of modeling a wide range of filters based on rank ordering such as the median. It was also capable of modeling linear finite impulse response **(**FIR) filters with positive weights. The framework was limited to modeling low pass filters or ‘smoothers’. More recently the framework was modified by [9]. This modification introduced the ability to model

both linear and nonlinear filters with negative as well as positive filter weights. It in effect opened up the possibility to model high pass and band pass filters as well as low pass filters.

Motivated by this success an image sharpening technique is developed and implemented through a framework of threshold decomposition. Consider an integer-valued set of samples x_1 , x_2 , … , x_n forming the signal X = (x_1 , x_2 , … , x_n) where $x_i \in$ {-m …, -1, 0, 1, … , m}. The threshold decomposition of X amounts to decomposing this signal into 2m binary signals X^{-m+1}, …, X^{0}, …, X^{m}, where the ith element of x^{m} is defined by the Eq. (2):

$$x_i^m = \begin{cases} 1 & \text{if } x_i \geq m \\ -1 & \text{if } x_i < m \end{cases} \qquad (2)$$

The above threshold decomposition is reversible, such that if a set of threshold signals is given, each of the samples in X can be exactly reconstructed as shown in Eq. (3):

$$x_i = \frac{1}{2} \sum_{j=-m+1}^{m} x_i^j \qquad (3)$$

Thus, an integer-valued discrete-time signal has a unique threshold signal representation, and vice versa.

2.3 Image Sharpening

The principle for image filtering method and edge detection can be done by several techniques. Firstly, signal reduction is required to emphasize the edge and brighten the image. In this case, high pass filter is used to filter the signal as well as to detect the edges from the original image. Hence, the solution for this process is the total of the original image and the edge as the Eq. (4):

$$f_s(x_i, y_i) = f(x_i, y_i) + \Omega F(f(x_i + y_i)) \qquad (4)$$

Where:
$f(x_i,y_i)$ = The original pixel value at the coordinate (x_i,y_i)
F(.) = The high pass filter
Ω = A tuning parameter which is greater or equal to zero
$f_s(x_i,y_i)$ = The sharpened pixel at the coordinate (x_i,y_i)

The value represents Ω as the perspective degree of sharpness, the higher the Ω the more sharpened is the image. Another well known technique which enhances blur images is called Unsharp Masking (UM) technique. The solution of this technique begins by subtracting the original image with the blur image. In the other words, subtract low pass filter from the input image. These results for the output image which emphasizes on the detail and sharpness [10]. Generally, blurred images occur by several low pass filtering in the image. Hereby was the Eq. (5) for Unsharp Masking technique:

$$f_s(x_i, y_i) = f(x_i, y_i) - f_b(x_i, y_i) \qquad (5)$$

Where:
$f_s(x,y)$ = The sharpened image obtained by unsharp masking
$f_b(x,y)$ = The blurred version of $f(x_i,y_i)$

According to this equation, increase in sharpness is eligible by using high boost filter [11]. The relations between the above two equations were as shown in Eq. (6)

$$f_{sh}(x_i, y_i) = A * f(x_i, y_i) - f_b(x_i, y_i)$$

or: (6)

$$f_{sh}(x_i, y_i) = (A - 1) + f_h(x_i, y_i)$$

A = A variable which is greater or equal to 1
$f_{sh}(x_i,y_i)$ = The high boost sharpened image
$f_b(x_i,y_i)$ = The low pass filter of $f(x_i,y_i)$
$f_h(x_i,y_i)$ = The high pass filter of $f(x_i,y_i)$

2.4 The Discrete Wavelet Transform

The generic form for a one-dimensional (1-D) wavelet transform is shown in Figure (1). Here a signal is passed through a lowpass and highpass filter, *h* and g, respectively, then down sampled by a factor of two, constituting one level of transform. Multiple levels or “scales” of the wavelet transform are made by repeating the filtering and decimation process on the lowpass branch outputs only. The process is typically carried out for a finite number of levels K, and the resulting coefficients, $d_{i1}(n)$, $i \in \{1,\dots,K\}$ and $d_{K0}(n)$, are called wavelet coefficients.

Only the maximally decimated form of the wavelet transform is used, where the downsampling factor in the decomposition and upsampling factor in the reconstruction equals the number of filters at each level (namely two).

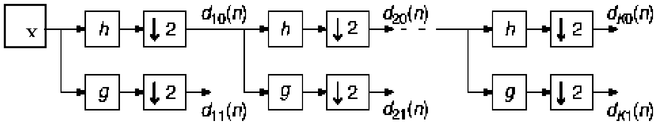


Figure 1: A K-level, 1-D wavelet decomposition. The coefficient notation d_{ij}(n) refers to the jth frequency band (0 for low and 1for high) of the ith level of the decomposition.

Referring to Figure (1), half of the output is obtained by filtering the input with filter H(z) and down-sampling by a factor of two, while the other half of the output is obtained by filtering the input with filter G(z) and down-sampling by a factor of two again. H(z) is a low pass filter, while filter G(z) is a high pass filter.

The 1-D wavelet transform can be extended to a two-dimensional (2-D) wavelet transform using separable wavelet filters. With separable filters the 2-D transform can be computed by applying a 1-D transform to all the rows of the input, and then repeating on all of the columns. Using the Lena image in Figure (2a) shows an example of a one-level *(K =1)*, 2-D wavelet transform, The example is repeated for a two-level *(K=2)* wavelet expansion in Figure (2b).

In two dimensions, usually apply filtering both horizontally and vertically. Filtering in one-direction results in decomposing the image into two “components”. The total numbers of produced “components” after the vertical and horizontal decompositions is four. These 4-components are referred to as image subbands, LL, HL, LH, HH. The first subband (the LL subband) will contain low pass information, which is essentially a low-resolution version of the image. Subband HL will contain low pass information vertically and high pass information horizontally, and subband LH will contain low pass information horizontally and high pass information vertically. Finally, subband HH will contain high pass information in both directions [13].

From Figure (2a) subband LL is more important than the other 3 subbands, as it represents a coarse version of the original image. The multiresolutional features of the wavelet transform have contributed to its popularity.

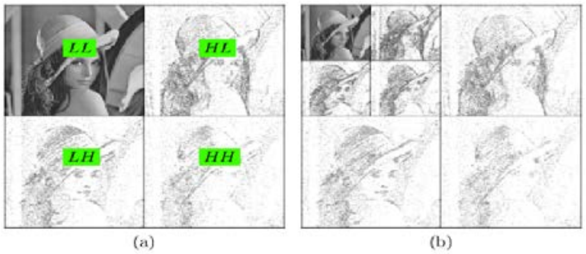


Figure 2: (a) One level wavelet transform in both directions of a 2D signal. (b) Two levels of wavelet transform in both directions.

3. The Proposed Scheme

3.1 Sub-band Coding

Sub-band coding is a coding strategy that tries to isolate different characteristics of a signal in a way that collects the signal energy into few components. This is referred to as energy compaction. Energy compaction is desirable because it is easier to efficiently code these components than the signal itself [16].

The sub-band coding scheme tries to achieve energy compaction by filtering a signal with filters of different characteristics. By choosing two filters that are orthogonal to each other and decimating the output of these filters a new two component representation is achieved as shown in Figure (3). In this new representation, hopefully, most of the signal energy will be located in either **a** or **d**.

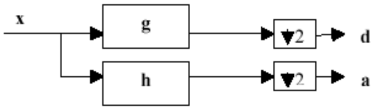


Figure 3: Splitting of the signal x into two parts.

The filters **h** and **g** are usually low-pass and high-pass filters as mentioned previously. The two components **a** and **d** will then be a low-pass and a high-pass version of the signal x. Images have a typical low-pass characteristics, and this is the reason why we should expect **a** to contain most of the energy if **x** is an image.

Besides trying to achieve energy compaction the filters **h** and **g** should be chosen so that perfect reconstruction of **x** from **a** and **d** is possible. In Figure (3) a two-component representation of **x** is achieved. It might be desirable to divide the signal into more components. A more common choice however is to cascade the structure in Figure (3). There are two major strategies for cascading the filters, the hierarchical structure and the flat structure. In the hierarchical structure the output from the low-pass filter is treated as the input to a new filter pair as depicted in Figure (4). While in the flat structure both the low-pass and the high-pass outputs are inputs to a filter pair, this structure is depicted in Figure (5). In both figures the corresponding splitting of the frequency axis is also shown. The process of dividing the signal into components will be referred to as decomposition or transform.

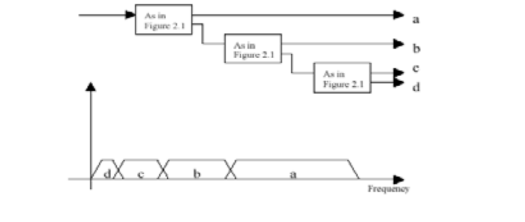


Figure 4: The hierarchical filter structure

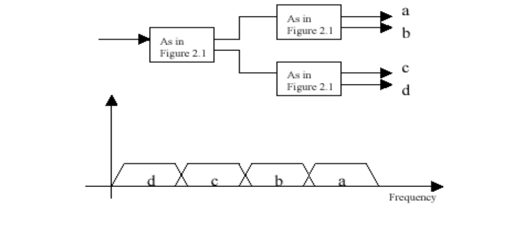


Figure 5: The flat filter structure

3.2 Extension to More Dimensions

To be able to use sub-band coding for images, the scheme above has to be adapted to two-dimensional signals. The extension of the sub-band coding scheme to higher dimension is straight-forward. Apply the filters repeatedly to successive dimensions. For a NxN image, first compute N one-dimensional transforms corresponding to transforming each row of the image as an individual one-dimensional signal. This will result in 2 NxM sub-images, one corresponding to the low-pass filtered rows and one corresponding to the high-pass filtered rows. Each of these sub-images is then filtered along the columns splitting the data into 4 MxM sub-images (low-pass row low-pass column, low-pass row high-pass column, high-pass row low-pass column, high-pass row high-pass column). This completes the one stage of the decomposition of an image. The process is depicted in Figure (6).

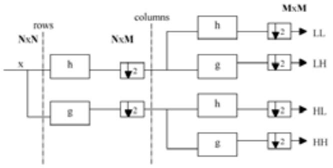


Figure 6: One stage of a two-dimensional decomposition

3.3 Wavelet Packets

The concept of wavelet packets (WP) extends the octave-band tree structured filter bank to consist of all possible frequency splits, so that best decomposition topology can be chosen to suit the individuality of different images. Figure (7) shows all of the possible representations of a wavelet packet decomposition of maximum depth two, among which the full tree is at the first position from left.

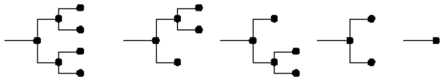
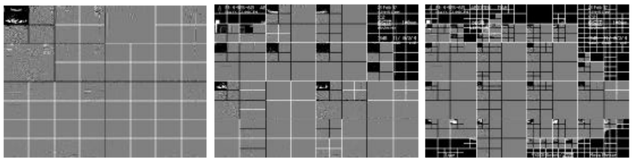


Figure 7: Possible wavelet packet trees, maximum depth=2

The best basis subtree is chosen among the entire library of wavelet packet bases. The problem of bit allocation for the decomposed subbands at a target bit rate is concerned with a set of given admissible quantization choices.

Each decomposition divides an image into four quadrants. Two-dimensional frequency partition produces four subbands: LL, LH, HL and HH. General space-frequency segmentation applies the general time-frequency-pruning algorithm to choose between the four-way splits in space or frequency in a space-frequency tree. This algorithm should generate a better optimal basis than both the single-tree and the double-tree. Its basis must be at least as good as the best double-tree/single-tree basis, because the set of possible double-tree/single-tree bases is a subset of the possible SFS bases.

An example for this partition is shown in Figure (8). It has maximum decomposition depth of five. The white lines indicate that the sub-image is space split, while the black lines indicate that frequency segmentation is applied. If there is no line inside the sub-image, this will indicate that there is no splitting performed on the sub-image.



level=3 level=4 level=5

Figure 8: partitions

The adaptive partitioning is done as follows:

1. Decompose an image component into blocks of fixed size (say 128 or 64).
2. If the size of the considered subblock is greater than MinSiz then this subblock must be tested to determine whether it requires further partitioning or not.
3. Initial subblocks will be partitioned, either as spatial (space) partitioning or as frequency (DWT) partitioning.
4. The Median Adaptive Predictor (MED) is utilized to evaluate whether the partitioning is performed in space or frequency.
5. There is one parameters have a great effects on selecting an appropriate partitioning types (the threshold). This operation is performed after doing the space and frequency partitioning on the same block, and then selecting partitioning type whose

error is less than other and it is higher than a selected threshold.

6. According to these conditions, the partitioning operation is done on all the image block. These operations continue until the block size reaches the MinSiz or they are uniform and the error of MED predictor doesn't exceed the threshold value.
7. the algorithm of partitioning as follows

```
1)initialization: /* Wid & Hgt are the width
  and height of image */
    Nx=Wid div MaxSiz; Ny=Hgt div MaxSiz; L=0;
    Input Rat and Thr by the user
  2) for each (Iy,Ix) where Iy=0 to Ny-1 :
  Ys=Iy*MaxSiz; and Ix=0 to Nx-1: Xs=Ix*MaxSiz;
  do:
       2.1) read all the fields of initialized
  partitions:
            Part[L].X=Xs; Part[L].Y=Ys;
  Part[L].Siz=MaxSiz;
  Part[L].Typ=0; Part[L].Nxt=L+1;
       2.2) increment L by 1
    end for each
3) put –1 at the last field Nxt of the last
  partition
4) compute the minimum size:
  MinSiz=MaxSiz/$2^{levels}$;
5) initialize J=0;
6) repeat the flowing operations:
      if Part[J].Siz>MinSiz then put the
  coefficients in A[]
6.1) Send A[] to the MED of space with its size
  and return the Error E1
         6.2) Apply wavelet transformation A[]
  and return B[]
6.3) Send B[] to the MED of frequency with its
  size and return the Error E2
6.4) if (E1>Thr) or (E2>Thr)
if (E1<=Rat*E2) then partition the block with
  space segmentation
else partition the block after applying DWT
      else J=part[j].nxt
if J not equal -1 then goto step 6
      else exit
```

The algorithm of smoothing is as follows:

- A block (space partitioning) which is homogeneous is to be filtered by the moving of mean filter.
- A block (frequency partitioning) which has relative weak edges or un-continued edges, are filtered by the mode filter.
- A block (reminder) which has sharp edges is to be filtered by the median filter.

The following figure shows the proposed scheme:

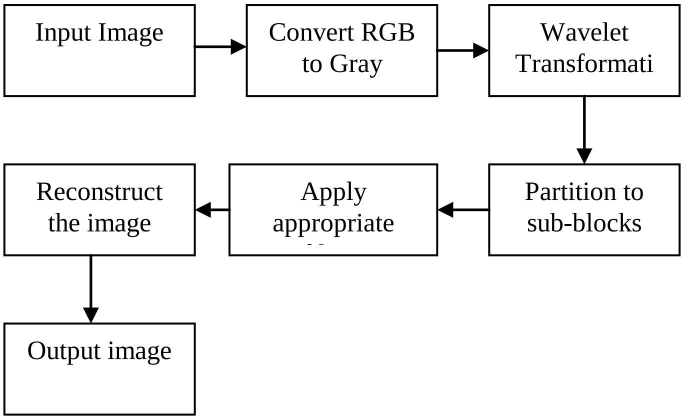


Figure 9: The Proposed Scheme

3.4 Image quality evaluation

The result image can be evaluated with two characteristics, distortion and sharpness. According to the distortion evaluation, adjusting errors are required, by computing the Mean Square Error (MSE). Mean square error has been the performance metric in lost performance. Peak Signal to Noise Ratio (PSNR) adjusts the quality of the image which the higher the PSNR refers to the better quality is the image [11]. The formula for MSE and PSNR are Eq. (7) and Eq. (8)

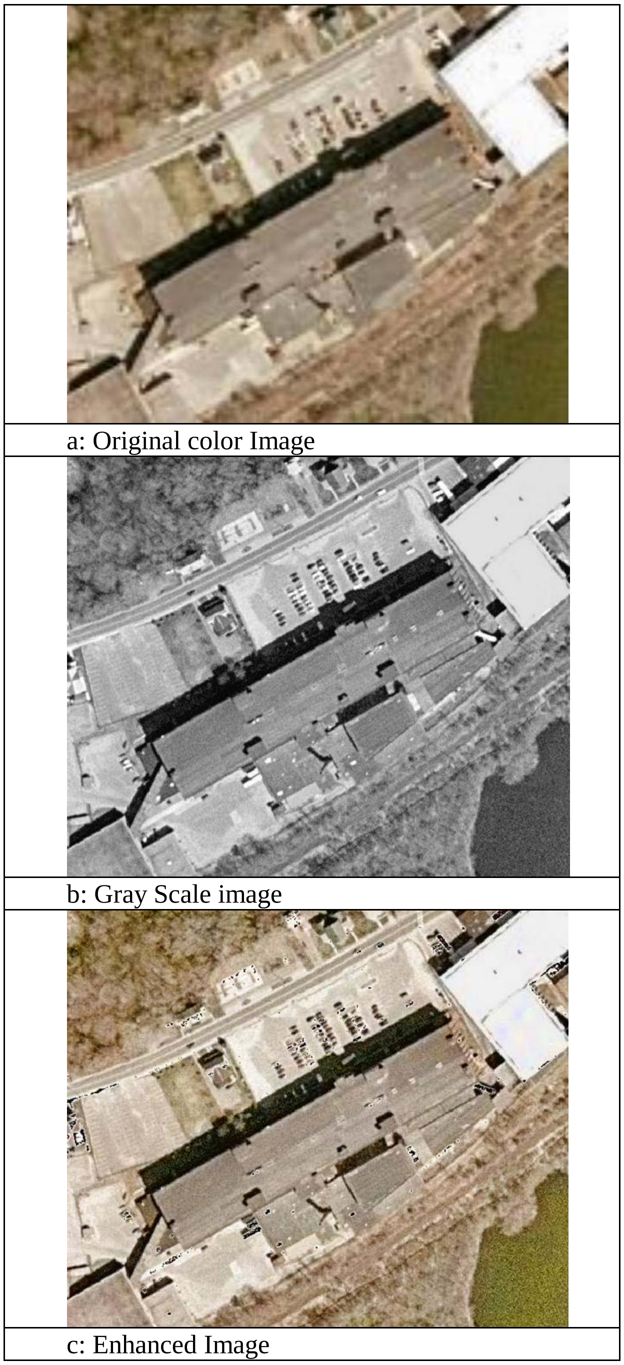
$$MSE = \frac{\sum_{MN}[I_1(m,n)-I_2(m,n)]^2}{M.N} \qquad (7)$$

$$PSNR = 10\log_{10}\left[\frac{R^2}{MSE}\right] \qquad (8)$$

The MSE expression is generally referred to the absolute error equation because the former error is analytically tractable. The most common error in image processing is the normalized brightness of the image. In the previous equation, M and N are the number of rows and columns of the input image, respectively. Then, all the blocks would compute the PSNR. In the PSNR equation, R is the maximum fluctuation in the input image data type.

4. Results

This section presents application results for the enhancement of satellite images. The images tested in the research were performed shown in Figure 10 and 11which was express in the numerical form of satellite image.

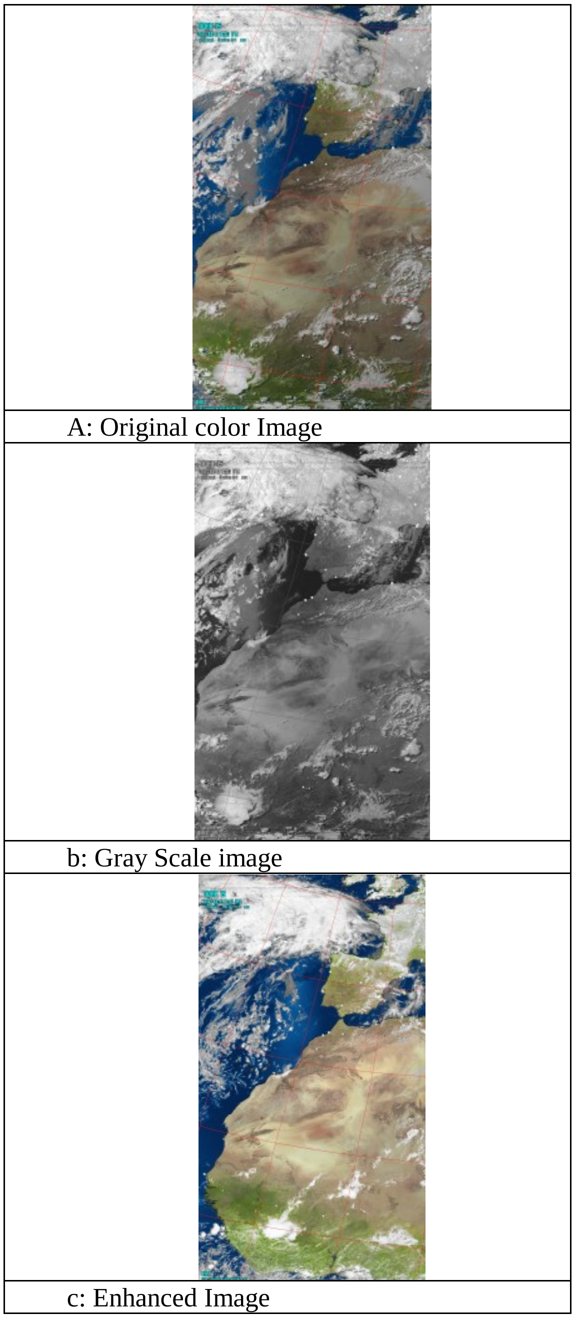


a: Original color Image

b: Gray Scale image

c: Enhanced Image

Figure 10: The Enhanced Image



A: Original color Image

b: Gray Scale image

c: Enhanced Image

Figure 11: The Enhanced Image

In order to compare the results accurately, Figure 10 and 11 represents the final step which includes subtracting the edge detected image from the original image. As a consequence, the MSE of the output image and the input image values at 0.63 and 0.49, respectively. In this study, we determined the total difference by comparing the PSNR between the original image and the output image, which the PSNR values at 25.67 and 28.95 dB, respectively.

Practically, the PSNR which values higher than 30 dB is invisible for human sight to analyze the color distortion between two images. The experimental result with the new algorithm procedure shown in figure10 and 11. Figure (a) is the original image and figure (b) is the result of color to grayscale transformation and finally, figure (c) show the result of the enhanced.

5. Conclusions

This paper introduced a new enhancement filter for digital satellite images. In the proposed scheme, the edge detected guided smoothing filters succeeded in enhancing low satellite images. This was done by accurately detecting the positions of the edges through threshold decomposition. The detected edges were then sharpened by applying smoothing filter. By utilizing the detected edges, the scheme was capable to effectively sharpening fine details whilst retaining image integrity. The visual examples shown have demonstrated that the proposed method was significantly better than many other well-known sharpener-type filters in respect of edge and fine detail restoration.

6. References

[1] Nick Kanopoulos, Nagesh Vasanthvada and Robert L Baker, “Design of an Image Edge Detection Filter Using the Sobel Operator”, IEEE Journal of Solid-State Circuits, (1988), Vol. 23, No. 2, pp. 358-367.

[2] Bin Wang D, Rose M and Aly A Farag, “Local Estimation of Gaussian-Based Edge Enhancement Filters Using Fourier Analysis”, IEEE Transactions on Acoustics, Speech, and Signal Processing, (1993), Vol. 5, pp. 13-16.

[3] Day-Fann Shen, Chui-Wen Chiu and Pon-Jay Huang, “Modified Laplacian Filter and Intensity Correction Technique for Image Resolution Enhancement”, IEEE International Conference on Multimedia and Expo, (2006), Vol. 7, Nos. 9-12, pp. 457-460.

[4] Cheevasuvit F, Dejhan K and Somboonkaew A “Edge Enhancement Using Transform of Subtracted Smoothing Image”, ACRS, (1992), Vol. 3, No. 12, pp. 23-28

[5] Jin Jesse S “An Adaptive Algorithm for Edge Detection”, MVA’SO IAPR Workshop on Machine Vision Applications, (1990), Vol. 9, November 28-30, pp. 14-17.

[6] J. M. Prewitt, 1970, "Object enhancement and extraction," Picture Processing and Psychopictorics, pp. 75-149.

[7] D.Stark and W.Bradley Jr., Ed. 1992, Magnetic Resonance Imaging, St. Louis, MO: Mosby.

[8] J. P. Fitch, E. J. Coyle, and N. C. Gallagher, "Median filtering by threshold decomposition," IEEE Transactions on Acoustics, Speech and Signal Processing, 1984, vol. 32, pp. 1183-1188.

[9] G. R. Arce, "A general weighted median filter structure admitting negative weights," IEEE Transactions on Signal Processing, 1998, vol. 46, pp. 3195-3205.

[10] Xu, D. and R. Wang, 2009. An improved FoE model for image deblurring. Int. J. Comput. Vis., 81: 167-171. DOI: 10.1007/s11263-008-0155-3

[11] Gonzales, R.C. and R.E. Woods, 2002. Digital Image Processing. 2nd Edn., Prentice Hall, USA., ISBN: 10: 0130946508, pp: 793.

[12] Chen, Z.Y., B.R. Abidi, D.L. Page and M.A. Abidi, 2006. Gray Level Grouping (GLG): An automatic method for optimized image contrast enhancement-Part I: The basic method. IEEE Trans. Image Process., 15: 2290-2302. DOI: 10.1109/TIP.2006.875204

[13] C.A. Christopoulos, T. Ebrahimi and A.N. Skodras. “JPEG2000: The New Still Picture Compression Standard”, Media Lab, Ericsson Research, Ericsson Radio Systems AB, S-16480 Stockholm, Sweden. http://www.eecs.harvard.edu/~michaelm/CS222/jpegb.pdf

Dr. Muna F. Al-sammarai. She obtained her B.Sc. degree from Al-Mansour University Baghdad-Iraq in 1992, M.Sc. form University of Baghdad and PhD degrees from University of Technology, Baghdad-Iraq in 1997 and 2002 respectively. She is an Assistant Prof. at MIS Dept., Faculty of Faculty of Economics and Administrative Sciences in Al-Zaytoonah University, Amman, Jordan. Her research interests include: Software Engineering, Image Processing, and Intelligent Systems.

Dr. Nedhal A. Al-Saiyd. She got her B.Sc. degree from University of Mosul-Iraq in 1981, M.Sc. and PhD degrees from University of Technology, Baghdad-Iraq in 1989 and 2000 respectively. She is an Assistant Prof. at Computer Science Dept., Faculty of Information Technology, in the Applied Science University, Amman, Jordan. Her research interests include: Software Engineering, Ontology Engineering, Intelligent Systems, User Authentication, e-learning and Speech Processing.

A Fuzzy Realistic Mobility Model for VANET

Alireza Amirshahi[1], Mahmood Fathi[2], Morteza Romoozi[3] and Mohammad Assarian[4]

[1] Computer Engineering Department, Islamic Azad University, Arak Branch
Arak, Iran

[2] Computer Engineering Department, Science and Technology University, Tehran Branch
Tehran, Iran

[3,4] Computer Engineering Department, Islamic Azad University, Kashan Branch
Kashan, Iran

Abstract

Realistic mobility models can assess more the results more accurate estimate parameters because it is closer to the real world. In this paper a realistic Fuzzy Mobility Model has been proposed. This model has rules which are changeable depending on nodes and environmental conditions. This model is more complete and precise than the other mobility models. After simulation, it was found out that not only considering nodes movement as being imprecise (fuzzy) has a positive effects on most of ad hoc network parameters, but also, more importantly as they are closer to the real world condition, they can have a more positive effect on the implementation of ad hoc network protocols.

Keywords: *Mobility Model, Ad hoc Networks, Realistic Mobility Model, Fuzzy Systems, Nodes Signal*

1. Introduction

Nowadays ad hoc networks have been used in a variety of applications. Mobility models in ad hoc networks are of special importance. Mobility model identifies the primary place of nodes and the manner of nodes mobility. Mobility models fall into two categories: realistic and unrealistic. As realistic mobility models are more similar to real world conditions, they provide more accurate results.

Vehicular ad-hoc networks (VANETs) are a particular kind of mobile ad-hoc networks where nodes are embedded into Moving vehicles, equipped with short-range wireless communication devices and positioning systems like GPS or Galileo. VANETs are gaining popularity in both academia and Industry as a key technology for many emerging services and applications in the automotive field, e.g. safety, traffic optimization and infotainment.

Before applying to the real world, computer simulation is a valuable tool for evaluating protocols and other network parameters. Simulation can be applied easily, while implementation of ad hoc networks in the real world is difficult and expensive. Moreover, simulation has other advantages such as iterative scenario, parameter isolation, and measuring different metrics. Glomosim [1] and NS2 [2] are the most famous simulators used for evaluating and comparing computer network protocols. Mobility model, signal propagation model and routing protocol are the most important parts of wireless simulators. Many realistic models have been presented in most of which nodes mobility is random and simulation environment is free, without obstacle and pathway.

Applying these models cannot represent the efficiency of the networks in real condition because in real condition, nodes must move in predefined passages and nodes signals must be blocked by obstacles. The movement patterns and path selections are not random. Some realistic models have been presented so far like Graph-based Mobility Model [3] and Obstacle Model [4] and. In these models, there are usually obstacles and pathway, but no attention has been paid to the movement patterns and destination selection. Meanwhile, the type of destination selection of nodes is not random. For instance, the selection of people’s destination in a VANET environment is not random and many parameters are involved such as time, current place, the priority of going to different places and etc.

The mobility of a mobile node and its mobility environment are not precise. Namely, a urban environment is not precise because every place has different parts and precise coordination of each part cannot be stated.

As fuzzy control systems are capable of solving imprecise problems efficiently, by using fuzzy control system in the proposed Fuzzy Mobility Model, the motion rules of different kinds of nodes, based on type of the activity and environment, have been designed. Fuzzy control system includes fuzzy rules which describe the nodes mobility in an adaptable way with the environment. This model has a knowledge base which can be changed based on nodes conditions, types of nodes and environment. By using such knowledge base, the mobility rules of every environment can be imposed upon a mobility model as an input, until the mobility is created in that specific environment.

A review of the related studies has been presented in part 2. Part 3 contains the proposed Fuzzy Mobility Model. The simulator (Glomosim[5]) and its results have been presented in part 4 and the conclusion has been mentioned in part 5.

2. A Review of The Related Studies

Regarding realistic mobility models, many studies have been done, but most of them have been performed on environment model and signal blockage and just a few attention has been paid to real movement patterns. In these models, destination selection of nodes was either completely random and the selection of path by algorithm was either the shortest one or it was selected randomly which are not considered suitable. For instance, the Obstacle Mobility Model [4], presented in 2003 by A. Jardesh, is one of the most successful realistic models. This model has an appropriate signal blockage and environment sub-model and it can be used as signal blockage and environment sub-model of other models, but it lacks a real world-based mobility pattern model. The destination selection of this model is entirely random and the path selection is done by Dijkstra algorithm with the shortest path regarding the number of edges which is not a suitable criterion.

A realistic group model, called OCGM [6], based on obstacle and mobility model RPGM [7] has been proposed which has similar environment model and blockage signal, but in its movement pattern sub-model, nodes move in groups.

Graph-based [3] is another model, environment model of which is constituted by a graph and this graph is the paths of a map and has not a specific signal blockage sub-model.

The next model which is based on Graph-based Model, named Area Graph-based [8], has been represented and its environment sub-model is similar to Graph-base Model and lacks signal blockage model, but compared to Graph-based Model, its movement pattern sub-model has been improved. In the other words as long as the nodes are inside the graph vertices, they have Random Waypoint [9] mobility. But for leaving vertex, nodes must select one of the output edges of the vertex which has probability from the beginning of simulation, along with related probability.

Still another realistic model, called Environment Aware Mobility [10] has been represented. The environment sub-model of this model is different from that of Graph-based Model mentioned above. In this mobility model, the environment is divided into a series of sub-environment inside of which there are some obstacles and movement pattern sub-model of nodes in each sub-environment can be one of the random mobility models. This model has signal blockage sub-model.

There are some realistic mobility models which touch on nodes movement pattern models. But the number of these models is by far fewer than the other models. For example, we can refer to a Cluster-based Mobility Model [11] for intelligent nodes by M.Romoozi. This mobility model has focused on the movement pattern sub-model and has improved it.

H. Babaei has proposed another model in 2007, named Obstacle Mobility Model Based on Activity Area [12]. This model has used environment sub-model and signal obstruction model of obstacle, but in this model the node movement pattern has been improved and a different range of activity and speed has been assigned to each group of nodes. Nodes select those vertices which are closer to the area of the activity with greater probability, but this selection is done by Dijkstra algorithm.

There are other models such as Manhattan Mobility Model [13], Free Way Mobility [13], and Urban Mobility Model [14]. But compared to the more complete models mentioned above, these models are of little importance.

2.1 Classification of Mobility Models

Mobility models can be divided into two categories: realistic and unrealistic. In realistic mobility models, the mobility of nodes is assessed in the real world conditions. In this model, not only mobility pattern of mobile nodes is considered, but also simulation environment and the effect of environment on signal nodes are examined. In unrealistic mobility models, a free and without obstacle space is taken into account in which nodes move freely everywhere and their selection of destination and path is usually random and there is no predefined obstacle and pathway for them. These models do not determine an accurate result in evaluating ad hoc network protocols.

2.1.1 Realistic Mobility Models

Realistic mobility models create an environment similar to the real world. This environment includes some pathways through which nodes must move in these pathways. This model also includes some obstacles. Not only these obstacles obstruct the nodes movement, but also they weaken or remove nodes signal. The more the environment is similar to the real world conditions, the more accurate the evaluation results will be.

Considering the analyses have done up to now, each realistic model is usually composed of 3 sub-models. These sub-models are related to one another and they can hardly be separated as follow:

-- Environment sub-model.
-- Signal obstruction sub-model.
-- Movement pattern sub-model.

Environment sub-model includes environmental obstacles such as buildings, mountains and etc., which usually exist in real environment. This sub-model also includes some paths that exist among these obstacles. These paths force the nodes to move only through these paths.

Signal obstruction sub-model in the proposed mobility model does not include obstacles.

Movement pattern sub-model includes the manner of destination selection, the selection of path toward the destination, and the amount of pause in destination. To sum up, the manner nodes movement is explained in environment sub-model.

2.2 Fuzzy Control System

The term ‘fuzzy’ means imprecise. Although fuzzy systems describe uncertain and unclear phenomena, Fuzzy theory is a precise one. The heart of a fuzzy system is a knowledge base which is composed of fuzzy If-Then rules.

The main structure of fuzzy systems is shown in figure 1.

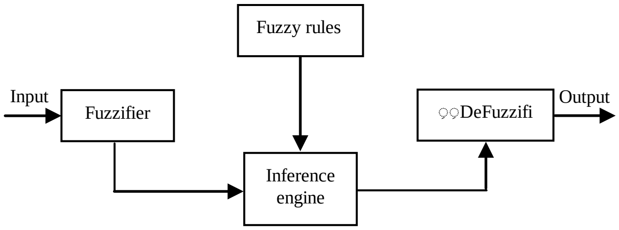


Fig. 1 The main structure of fuzzy systems.

The fuzzifier used in Fuzzy Mobility Model is a unique fuzzifier (1). This fuzzifier maps a singular point x*∈u with real value on a fuzzy unique A′ in u and the membership function in X* equals one and in other points u equals 0. It means:

$$u_{A'}(x) = \begin{cases} 1 : x = x^{*} \\ 0 : \text{O.W} \end{cases} \tag{1}$$

Defuzzifier used in Fuzzy Mobility Model is the center average defuzzifier. Center average defuzzifier is the most commonly used defuzzifier in fuzzy systems and fuzzy control systems. Fiscally it is simple, and at the same time, intuitively it is justifiable. Center average defuzzifier can be defined in the following way:

$$y^{*} = \frac{\sum_{\ell=1}^{M} y^{-\ell} w_{\ell}}{\sum_{\ell=1}^{M} w_{\ell}} \tag{2}$$

In this equation (2) $y^{-\ell}$ is the center of ℓ fuzzy set and w_{ℓ} is its hight degree and M is the number of our rules.

Inference engine of used in the fuzzy mobility model is multiple inference engine (3).

$$u_{B'}(y) = \max_{\ell=1}^{M}\left[\sup_{x\in u}(u_{A'}(x)\prod_{i=1}^{n} u_{A_i}\ell(x_i)u_{B}\ell(y))\right] \tag{3}$$

3. The Proposed Fuzzy Mobility Model

In a real environment, nodes are divided into different groups with similar mobility features. For instance, in VANET we have personal automobile nodes, public automobile nodes, and ambulance automobile nodes. For each group of nodes, the manner of destination selection, the movement speed, time and etc., are different.

In the real world, the destination of nodes is expressed imprecisely and the conception of fuzziness is hidden in it. For example, emergence place is included different sections, and we cannot express a precise coordinates. For example, the emergence place cannot be stated in unique X and Y points.

In the real world, nodes destination is selected based on the time. Namely, in VANET, a personal automobile goes to a university in the morning and to the residential complex at noon, but these times are not stated exactly. Some people believe that morning starts from 7 to 10, but others believe it to be from 7 to 9 and etc. So time can be considered as being fuzzy.

Regarding that each of the nodes has a different mobility, so the destination selection of each node will be different from the others. To provide an example, in a VANET environment, the mobility of a personal automobile node is different from that of a public or an ambulance. Therefore, their destinations are different.

3.1 Sub-Models of Fuzzy Mobility Model

Sub-models of Fuzzy Mobility Model include 3 parts:

3.1.1 Environment Sub-Model

In environment sub-model of the proposed mobility model is used a city map that Pathways in this sub-model are like a real environment. In fact we create a real environment to have a real simulated environment. In figure 2, buildings have been created like Squares and available edges in that are some pathways.

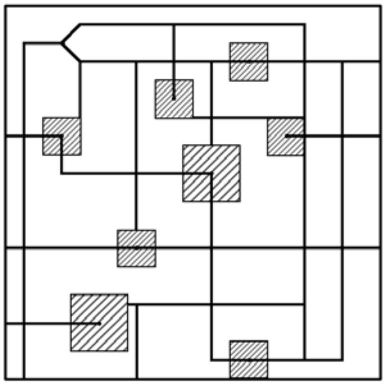


Fig. 2 Simulation environment.

3.1.2 Signal Obstruction Sub-Model

Whereas of the mobility model proposed there are no obstacles, in this mobility model, Signal obstruction sub-model does not use.

3.1.3 Movement Pattern Sub-model

The manner of movement, including the path selection, destination selection, and the amount of pause in destination, will be examined in this sub-model.

In this paper, the main focus is on the movement pattern of nodes. In the proposed Fuzzy Mobility Model, firstly, nodes are divided into groups with equal mobility features. Then, the manner of destination selection of nodes is defined by using a fuzzy control system (fuzzifier, fuzzy rules, inference engine and defuzzifier).This mobility model is suitable for mobility of a mobile node which has an imprecise mobility. The pass selection method in proposed mobility model is Dijkstra shortest pass algorithm.

The proposed mobility model gains from a fuzzy control system that contains fuzzy rules. These fuzzy rules describe the node mobility in an adaptable way to the environment. Fuzzy rules express the manner of destination selection for each group of nodes. As time and place inputs are expressed imprecisely (fuzzy), it seems that Fuzzy Mobility Model is more similar to the real world than the previous realistic model and this model can be used as a part of a simulator by MANET network researchers.

In the proposed Fuzzy Mobility Model, considering fuzzy systems (figure.1), fuzzifier input is time and place which are expressed precisely (for example for ambulance automobile node, the current place is hospital and the time is 8 o’clock). Fuzifier changes these amounts from being precise into fuzzy state and they go to the inference engine along with current fuzzy rules. Afterwards, the output of the inference engine goes to the defuzzifier and the next destination is defined based on the fuzzy rules.

3.2 Nodes Clustering

In VENET environment nodes can be divided into 3 groups: personal, public and ambulance. Each of these nodes has different motilities which are explained later on.

3.3 Mobility Analysis

Mobility analysis has different methods including locating the camera in specific places and identifying the movement of people and obtaining the related mobility model and the other method is using Radio-Frequency Identification (RFID). Still the next method is using questionnaire which has been in this mobility model. We distributed questionnaires among some VANET nodes and asked them to fill out the forms. This way we were able to identify the destination of these nodes in different time intervals. Because each node has different programs, the fuzziness of mobility model is proved.

3.4 Inputs of Fuzzy Mobility Model (VANET environment)

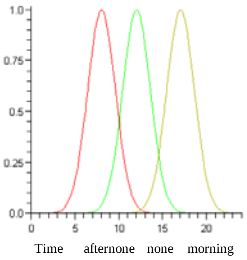


Fig. 3 Time input in Fuzzy Mobility Model.

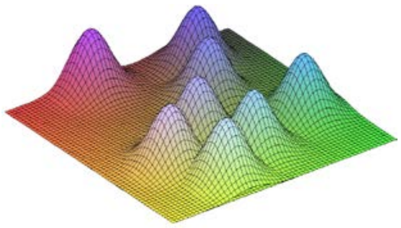


Fig. 4 Places of Fuzzy Mobility Model.

3.4.1 Time Input

Time input is divided into 3 parts: morning, noon and evening (according to figure. 3 mapped by Maple Software). A Gausian diagram has been applied to show the time as being fuzzy and its membership function has been shown in (4). A is a point in which the diagram has the highest value '1'. For instance, the value of a in the morning, at noon, and in the evening can be 8, 12, and 17 respectively.

$$\mu_{time}^{(t)} = e^{(-0.2(x-a)^2)} \qquad (4)$$

3.4.2 Place Input

The other input is place which is mapped by maple software (figure. 4). In order to show the place as being fuzzy, a Gausian diagram has been used and its membership function has been shown in (5). a and b are the coordinates of the center of places and their diagram in that points has the highest value. Coordinates of the center of sites has been shown in figure. 5.

$$\mu_{pos}^{(x,y)} = e^{(-(10^{-4}((x-a)^2+(y-b)^2)))} \qquad (5)$$

3.5 The output of the Fuzzy Mobility Model

The output of the Fuzzy Mobility Model is the selection of destination. Considering figure. 1(the main structure of Fuzzy Systems), the inputs of mobility model are the current time and place given to the fuzzy control system (given precisely). Now regarding the given rules table, the place of destination is defined.

3.6 Extracting the Rules Table

As mentioned before, VANET nodes have different mobilities in different times. So questionnaires have been used. Table 1 shows a type of the forms used for each of the nodes.

Table 1: Questionnaire forms of VANET nodes.

Question Form	Public automobiles □	personal automobiles □	Ambulance automobiles □

Morning		Noon		Evening	
Place	Priority	Place	Priority	Place	Priority
Hospital		Hospital		Hospital	
Emergency		Emergency		Emergency	
University		University		University	
Residential		Residential		Residential	
-Complex		-Complex		-Complex	
Park		Park		Park	
Bazaar		Bazaar		Bazaar	
City center 1		City center 1		City center 1	
City center 2		City center 2		City center 2	

Questionnaire forms were submitted to 70 personal automobiles, 70 public automobile and 70 ambulance automobile and they were all asked to read and fill out the forms carefully. For example, the ambulance automobile has to define the priority of going to the class of hospital by a digit between 0 and 1 for the morning, noon, and the evening in the questionnaire. He was asked to do the same for other places in the form. When the nodes returned the questionnaire, the average of priority of each node in the morning, at noon, and in the evening was calculated and this knowledge was used for problem solving.

In order to fill out the rules table, a map of the real urban environment is provided. Then, the centers of sites are defined by exact X and Y in a coordinate axis in which X and y have the maximum value of 10000. In figure. 5, the precise center of VANET sites is shown. For instance emergence place is located in coordinates X= 7500 and Y=6500.

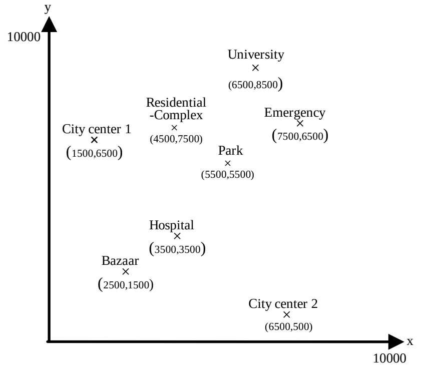


Fig. 5 Places coordinate of Fuzzy Mobility Model in VANET environment.

There are two important parameters in the proposed mobility model. These parameters are required for deciding the direction of nodes movement from one place to the other such as priority of nodes for moving towards the destination sites and distance of nodes from destination sites. In (6), both parameters have been taken into account. P_1 and P_2 are used for defining the weight of these parameters. This is a minimum equation.

$$\underset{Site=1}{\overset{n}{Min}}\, k = P_1(1-A)+P_2(\frac{d}{Max\_dist}) \qquad (6)$$

In this equation A is the priority of going to a place extracted from the questionnaire and d is the distance between the current place and node destination. As we have the coordinates of the center of sites, the distance between these two sites can be obtained from equation $d=\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$ (x_1 and y_1) are the coordinates related to the current site and (x_2 and y_2) are the coordinates of centers of the main destinations.(Current place and destination place are the sites shown in figure. 5). The distance between two sites (d) is divided by the maximum distance (Max_dis). According to this plan the farthest distance between two places equals 8060 (Max_dis=8060), so the results will be a digit between 0 and 1 (6). In conclusion, the more the result d/Max_dis is, the more the distance between two sites will be.

In (6), P_1 and P_2 are respectively the priority of going to the destination place and giving priority to the current place rather than destination place "$0 \le p_1,p_2 \le 1$". As the priority of going to a place is more important, the value of

P_1 is considered equal to 0.6. Regarding the result of d/Max_dis, instead of using A, (1-A) is used to create a balance in the equation. Now, the more the result of the equation p_1(1-A)+P_2(d/Max_dis), it points to the fact that the selection of this destination is not an appropriate choice .In each place we are, this equation must be repeated for each of other places (that can be one of the destination places) and finally the obtained figure, which has the least value, is selected as destination.

3.7 The Calculations Done in the VANET Environment

As we are in the current place and because of having 8 existing places figure. 5 in the current place, for destination selection, we should apply (6) 8 times and the selected destination will be the result in these 8 steps. Namely, in table 1 the node is personal automobile, so if the current place is the hospital and the time is morning, the destination place will be city center 2.
Tables 2, 3, and 4 show the rules of the personal, public, and ambulance nodes in Fuzzy Mobility Model.

Table 2: Personal automobile

place	Morning	Noon	Evening
Hospital	City center 2	City center 1	Bazaar
Emergency	City center 2	Residential-Complex	Park
University	University	Residential-Complex	Residential-Complex
Residential-Complex	City center 1	Residential-Complex	Park
Park	City center 1	City center 1	Park
Bazaar	Bazaar	City center 2	Bazaar
City center 1	City center 1	City center 1	City center 1
City center 2	City center 2	City center 2	City center 2

Table 3: Public automobile

place	Morning	Noon	Evening
Hospital	City center 1	Residential-Complex	Bazaar
Emergency	Residential-Complex	Emergency	Emergency
University	Residential-Complex	Residential-Complex	University
Residential-Complex	Residential-Complex	Residential-Complex	City center 1
Park	City center 1	Residential-Complex	City center 1
Bazaar	Bazaar	Bazaar	City center 2
City center 1	City center 1	City center 1	City center 1
City center 2	City center 2	City center 2	City center 2

Table 4: ambulance automobile

Place	Morning	Noon	Evening
Hospital	Hospital	Hospital	Hospital
Emergency	Emergency	Emergency	Emergency
University	Emergency	Emergency	Emergency
Residential-Complex	Emergency	Residential-Complex	Emergency
Park	Hospital	Emergency	Emergency
Bazaar	Hospital	Hospital	Hospital
City center 1	Hospital	City center 1	Hospital
City center 2	City center 2	City center 2	City center 2

It should be mentioned that obtaining rules table has different ways. We can seek help from experts, for example, to complete the rules table.

4. Simulation

The applied simulator is called Glomosim [5].

4.1 Simulation Parameters

The simulation environment is 10000 m × 10000 m and the least range for transfer of nodes is 250 m. Of course, because of the existence of the obstacles, the real transmission range of each node is limited. The propagation model is two-ray path loss. In MAC layers, IEEE 802.11 DCF protocol is applied and the band is 2mbps wide. As the nodes can be pedestrian and automobile we select the mobility speed of nodes between 0 m/s and 10 m/s. The stopping time will be selected randomly between 10 and 300 seconds. In different primary situations, each point of the diagram obtained from the average 20 time-simulation implementation with distributed nodes.
After the primary distribution of nodes in the vertices of Voronoi graph, nodes move for 60 seconds to be distributed all over the simulation environment. Then, 20 Data Session begins. The size of the data packet is 512 byte and the rate of transfer is 4 packets per second. The maximum number of packet which can be sent in each data session is 6000. So a heap of 6000 packet can be received by 20 destinations. Twenty sources and destinations are selected randomly. During the simulation, the movement continues for a period of 3600 second. All the data sessions apply CBR traffic model (a fixed bit rate). The numbers of clients and servers have been selected randomly.

4.2 The Manner of Fuzzy Mobility Model Application to Glomosim

The formula of the fuzzy systems [15] with inference engine of multiplication, unique fuzzifier (1) and center average defuzzifier (2) will be as follow (7):

$$f(x)=\frac{\sum_{l=1}^{M}\bar{y}^{l}\left(\prod_{i=1}^{n}\mu_{A_i^l}(x_i)\right)}{\sum_{l=1}^{M}\left(\prod_{i=1}^{n}\mu_{A_i^l}(x_i)\right)} \quad (7)$$

$x\in U\subset R^n$ Is the input of fuzzy system and $f(x)\in V\subset R$ is the output of fuzzy system. In Fuzzy Mobility Model the above mentioned formula is implemented in C language and then it is given to Glomosim simulator.

4.3 Evaluation Metrics

The main purpose of simulation is the examination and comparison of evaluation metrics. Fuzzy Mobility Model in the VANET environment is compared to other mobility models. Simulation has been done according to different speeds and now we examine the results. The evaluation metrics in the simulation done are as follow:

--Node Density: The average number of each node's neighbors is called node density.

--Broken Link Average: It is the average of broken links during the simulation.

--Average Data packet Reception: It is the number of receptions of the sent data packets in the desired destinations.

--Routing Overhead: It is the number of transfers of network layer controlling packets.

--End to End Delay: A delay which is required for a packet to arrive from the source to the destination.

The compared methods are as follow:

--FMM (Fuzzy Mobility Model).

--OMM(Obstacle Mobility Model).

--CBMM (Cluster Based-Mobility Model).

--RWMM(Random Waypoint Mobility Model).

4.4 The Results of Simulation of Fuzzy Mobility Model

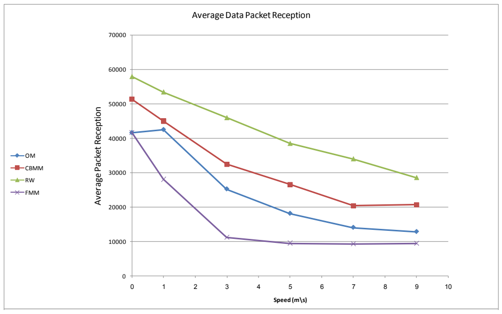


Fig. 5 Average data packet reception.

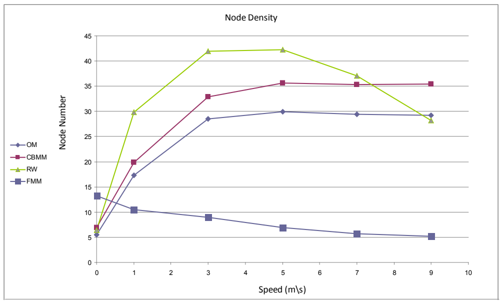


Fig. 6 Node density average.

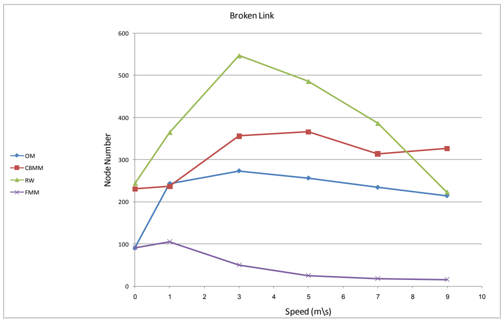


Fig. 7 Broken link average.

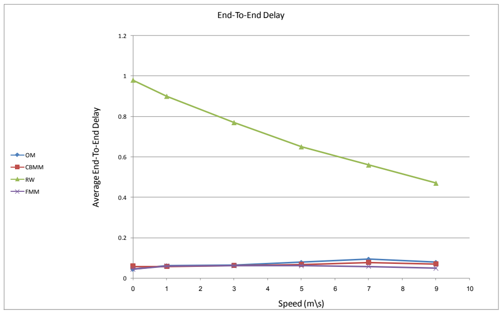


Fig. 8 End to End delay average.

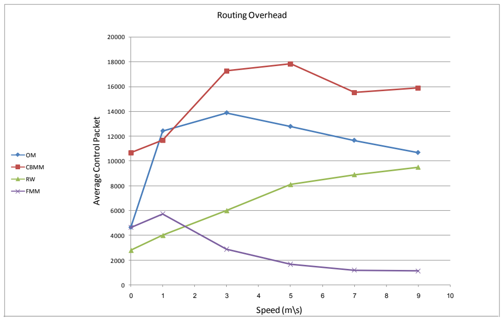


Fig. 9 Routing overhead average.

4.4.1 Average Data Packet Reception

Considering figure 5, it can be observed that Average Data Package Reception in RW Mobility Model is better than the other methods. In FMM Mobility Model Average Data Package Reception is lower, because nodes of the same type (for example, Public automobile node) are not beside one another. In CBMM Mobility Model, Average Data Package Reception is better, because nodes of the same type usually are beside one another.

4.4.2 Node Density

In RW mobility model node density is better other models, because all parameters (such as destination selection, route selection, and movement speed) are selected randomly. OM mobility model has node density of better rather than FMM mobility model, but the less node density has rather than two mobility models. Figure. 6 shows the average number of each node’s neighbors.

4.4.3 Broken Link Average

In FMM, pathways are stable, so fewer number of links break. RWMM has a greater number of broken links and generally OMM and CBMM have a fewer number of broken links than RWMM. It can be concluded that the more is the average number of each node’s neighbors in the related mobility model, the more broken links will be. Figure. 7 illustrates broken link average.

4.4.4 End to End Delay Average

In FMM mobility model, has the least end to end delay because pathways are stable. In both models (RW and FMM) there are no obstacles, but whereas the selection of destination and pathway in FMM mobility model against RW mobility model is not random, resulting end-to-end delay of that is lesser. End to end delay in two mobility model CBMM and OM is less than the RW mobility model. End to end delay average is illustrated in figure. 8.

4.4.5 Routing Overhead

From figure. 9, we can conclude that routing overhead in FMM mobility model is the least of other mobility models. But in other models due to pathways stable is less, the mobility models have to be sent more control packets. This makes the routing overhead is more than instead of proposed mobility model.

4.5 Conclusion

The focus of this paper is on the pattern of nodes movement in the real environment. The previous mobility models were either unrealistic not including obstacles and pathways, or realistic including obstacles and pathways similar to those of the real environment, but none of them pays any attention to the manner of nodes movement and destination selection.

Considering the examinations have done up to now, each realistic model is composed of 3 sub-models: Environmental sub-model, signal obstruction sub-model and movement pattern sub-model. There is a close relationship among these sub-models. In the proposed mobility model, environmental sub-model and the signal obstruction of obstacle mobility model have been applied, but it has a different movement pattern sub-model.

A fuzzy control system containing fuzzy rules has been used in this mobility model. In this paper, it has been proved that the mobility of a mobile node is fuzzy (imprecise) and also the mobility environment is a fuzzy one. The fuzzy control system used in this paper describes node mobility in an adaptable way to the environment. These rules describe the manner of destination selection. By using a fuzzy control system in the proposed Fuzzy Mobility Model, the movement rules of different types of nodes, depending on the kind of activity and environment and so on, have been imposed. This model also has a knowledge base which is changeable depending on nodes conditions, types of nodes and the environment. Using such knowledge base, movement rules of every environment can be imposed as input on the mobility model in order to consider the movement in that environment.

The type of mobility model, number, type of arrangement, size of obstacles and the speed of nodes movement are the parameters which have a considerable effect on the simulation results.

After simulation, it was found out that not only most of the results in Fuzzy mobility model have improved but also the nearest this model to real world conditions has helped to effectively. This model can help those researchers who would like to implement ad hoc networks protocols.

References

[1] L. Bajaj, M. Takai, R. Ahuja, K. Tang, R. Bagrodia, and M. Gerla, "Glomosim: A Scalable Network Simulation Environment:, Technical Report CSD, #990027, UCLA, 2003.

[2] The Network Simulator 2, http://www.isi.edu/nsnam/ns.

[3] j. Tian, J. Hahner, C. Becker, I. Stepanov, K. Rothermel, "Graph-based Mobility Model for Mobile Ad Hoc Network Simulation", in the Proceedings of 35th Annual Simulatin Symposium, in cooperation with the IEEE Computer Society and ACM. San Diego, California. 2002.

[4] A. P. Jardosh, E. M. Belding-Royer, K. C. Almeroth, and S. Suri,"Towards Realistic Mobility Models for Mobile Ad Hoc Netwotks", in Proceedings of ACM MOBICOM, San Diego, CA, 2003, pp. 217-229.

[5] M. Berg, M. Kreveld, M. Overmars, O. Schwarzkopf, "Computational Geometry: Algorithms and Applications", Springers Verlog, 2000.

[6] J. Kristoffersson, "Obstacle Constrained Group Mobility Model", in Department of Computer Science and Electrical Engineering Lulea University of Technology, Sweden, 2005.

[7] X. Hong, M. Geral, G. Pei, and C. C. Chaing, "The Performance of Query Control Schemes for the Zone Routing Protocol", in ACM SIGCOMM Describes ZRP Protocol, 1998.

[8] Bittner .Sven, Raffel .Wolf-Ulrich, and Scholz, “Manuel The Area Graph-based Mobility Model and its Impact on Data Dissemination Proceedings” of the 3rd Int’l Conf. on Pervasive Computing and Communications Workshops (PerCom 2005 Workshops) 2005.

[9] Q. Zheng, X. Hong, S. Ray, "Recent Advances in Mobility Modeling for Mobile Ad Hoc Network Research", in ACM-

SE 42 Proceedings of the 42th annual Southeast regional, Huntsville, Alabama, USA, 2004.

[10] Gang Lu, Belis Demetrios, Manson Gordon, “Study on Environment Mobility Models for Mobile Ad Hoc Network: hotspot Mobility Model and Route Mobility Model,” Wireless Com, Hawaii, USA, 2005.

[11] M. Romoozi, H. Babaei, M. Fathi, M. Romoozi, "A Cluster-Based Mobility Model for Intelligent Nodes", in LNCS., Verlag Berlin Heidelberg, 2009, pp. 565-579.

[12] H. Babaei, M. Fathi, M. Romoozi, "Obstacle Mobility Model Based on Activity Area in Ad Hoc Networks", in LNCS., Verlag Berlin Heidelberg, 2007, pp. 804-817.

[13] F. Bai, N. Sadagopan, A. Helmy, "The Important Framework For Analyzing The Impact of Mobility on Performance of Routing Protocols for Ad Hoc Networks", in Proceedings of IEEE INFOCOM, San Francisco, CA, 2003, pp. 825-832.

[14] S. Marinoni, H. Kari, "Ad Hoc Routing Protocol Performance in a Realistic Environment", in Proceeding of the 5th IEEE International Conference on Networking (ICN) , Mauritius, 2006.

[15] Wang, Lie-Xin, “A course in fuzzy systems and control.”

Alireza Amirshahi is currently Ms student at Islamic Azad University (Arak branch) in Iran. He was born in Kashan at 11 September 1973. He received a Bs in software engineering from the Islamic Azad University (Kashan branch) at 2002. He has taught in the areas of computer architecture and logic circuits and his research interests including computer architecture and Ad hoc networks.

Mahmood Fathy received his BSc in electronics from Iran University of Science and Technology in 1985, MSc in computer architecture in 1987 from Bradford University,
United Kingdom and PhD in image processing computer architecture in 1991 from UMIST, United Kingdom. Since 1991, he has been an associate professor in the Computer
Engineering School of IUST. His research interests include image and video processing, computer networks, including wireless and vehicular ad hoc network and video and image transmission over the Internet.

A Conceptual Framework of Knowledge Transfer in Malaysia E-Government IT Outsourcing: An Integration with Transactive Memory System (TMS)

Nor Aziati Abdul Hamid[1] and Juhana Salim[2]

[1] PhD Candidate, Information System Program, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600, Bangi, Selangor, Malaysia

[2] Professor, Information Science Program, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600, Bangi, Selangor, Malaysia

Abstract

Despite extensive research on knowledge transfer issues, there is a dearth of research that has explicitly focused on the role of Transactive Memory System (TMS) in enabling inter-organizational knowledge transfer in e-government IT outsourcing. Although the information systems literature has recently acknowledged the role of TMS in improving knowledge processes, most of the research is still remain in theoretical conjecture. Additionally, most of related research was done in the lab based on the physical, virtual or memory recall tasks. None of empirical work has been done in integrating TMS in outsourcing context since most researchers used interpretive approach. To address this gap, we applied positivist approach through operationalization of identified factors that give impact towards Malaysia Public Agencies outsourcing partnership. The present paper attempts to provide an integrated conceptual framework of knowledge transfer with and integration of TMS to facilitate knowledge transfer process which further can be validated.

Keywords: *Knowledge Management, Knowledge Transfer, Organizational Learning, Transactive Memory System (TMS), Information Technology Outsourcing (ITO)***.**

1. Introduction

Knowledge Management (KM) has been historically influenced by research undertaken across broad range of disciplines. These disciplines include sociology, psychology and philosophy. Until now, research in KM has been extended through various areas such as strategic management, information system, organizational learning, artificial intelligent and other more. Among those parent disciplines, organizational learning (OL) is the closets ‘cousin’ to knowledge management (KM). Hacket [1] considered KM and organizational learning as two sides of the coin. Transfer of knowledge is critical to knowledge-intensive project like IT outsourcing. However, the transfer of knowledge requires continuous organizational learning and the knowledge is being organized to enable knowledge retention capacity for future knowledge utilization. Knowledge is considered as tangible asset to organization. Tangible assets tend to depreciate in value when it is utilized. In contrast, knowledge grows substantially when it is fully utilized and depreciates or stagnant when it is not used. The organization needs to acquire the knowledge, learn, apply and reinvent the knowledge to make it suitable with the organization climate. Indeed, knowledge is of limited value if it is not shared and transferred throughout an organization. Thus, interest has increased in the phenomenon of how the firms create, retain, and transfer knowledge.

In the case of Malaysia, Malaysian Administrative and Modernization Planning Unit (MAMPU) [2] has created a "knowledge bank" structure in the public sector ICT framework to facilitate the sharing of knowledge and experience by capturing information across all Government agencies. This framework will create a structured and systematic transfer and utilization of knowledge generated. For the initial stage, several set of databases has been identified by MAMPU such as economic intelligence, security intelligence, R&D and Government statistics to create the knowledge bank. This initial project is implemented at four ministries; Finance, Health, Works and Education Thus, each ministry must develop their own knowledge bank with back end architecture that can integrate with other stated ministries. This project was initiated to address a high number of complaints regarding public services. There are many factors contribute to poor service delivery in the public sectors and one of them is low level of information and knowledge sharing among government agencies [3]. Although there is an increment in term of percentage complaints solved, the adoption and deployment of knowledge bank implementation reported by the scholars and how it can facilitate the knowledge transfer process is scarce. Furthermore, the knowledge bank focused more on the internal knowledge repositories among the public agencies without the absence of private agencies [2]. Riding on such issues, MAMPU formed a Special Interest Group for Knowledge Management (SIG-KMPS) in April 2010 to address the needs of KM in Malaysia Public Sector and to facilitate the public-private partnership. Since the government have been aggressively promoting the Shared Service and Outsourcing (SSO) industry, which undertake a full consideration of public-private partnership in supporting government transformation, it is crucial to consider a suitable framework of knowledge transfer that could support and facilitate transferring process during the outsourcing partnership.

Past researchers have suggested various organizational, human-related and IS-based mechanisms for improving knowledge transfer processes within and between organizations. Recent research has starts to integrate the concept of individual’s mental memory towards organization. Organization by itself is a combination of various stages of memories ranging from internal memories until external memories (e.g. stakeholders/shareholders). Therefore, this paper attempts to provide a better understanding of the phenomenon of knowledge transfer in IT outsourcing and how the transferring process may be bridged by applying organizational memory concept with existing identified factors during IT outsourcing project execution in Malaysia government setting. The study presented in this paper complements the existing knowledge transfer research in the context of e-government IT outsourcing project, while contributing to the body of empirical KM research. The reminder of this paper is structured as follow: next section will discuss research background and Malaysia public agencies e-government IT outsourcing in general. The following section discusses the relevant literature to develop underpinning theories. We proposed our conceptual framework of knowledge transfer in IT outsourcing in section four. We conclude the paper by some final remarks in section five.

2. Research Background

2.1 Knowledge Transfer and IT Outsourcing

Knowledge transfer (KT) is defined as a dyadic exchange between individuals, groups or organizations, in which a recipient can understand, learn and apply knowledge transmitted from a source [4],[5],[6]. A thorough review of literature reveals that many authors and researchers have failed to provide a clear cut definition for KT and at the same times use the term “knowledge sharing (KS)” and “knowledge transfer (KT)” interchangeably. However, recent scholars’ works have made a distinction line between these two terms. Knowledge sharing primarily concerned with the individual’s view while knowledge transfer concentrates more on the organizational view [7]. KS only takes the activities of giving or contributing, and is included under sub process of knowledge transfer. Furthermore, [8] asserts that KS does not include the receiving and reuse aspect of transfer. KT should involve active communication between two parties or active consultation for each other in order to learn what they both know. In a simple connotation, “people share knowledge” whereas “organizations transfer knowledge”.

Some researchers have been arguing of knowledge transfer concept since knowledge resides in employees (human components of organization), task and interrelationship, tools and technology (software and hardware) and network coordination (internal or external network coordination). There is no simplest way to transfer knowledge from a brain of a human to another brain perfectly and easily like transferring files form one computer to another. Hence, the nucleus of knowledge transfer process is the knowledge receiver. The knowledge receiver must have capability to learn, to understand and to know for applying in right circumstance. In line with that, all knowledge transfer mechanism incorporate social interaction either from direct interaction or virtual interaction. Ambos and Ambos [9] identified two mechanisms; (1) by personal coordination mechanism such as personnel motion, training, jobs rotation [10], interactions with suppliers and customers [11], community of practices and post-project reviews [12], (2) by technology based coordination mechanism such as collaboration software, distributed learning and business intelligence system. Most of Malaysia public organizations are actually practicing knowledge transfer using mechanism like staff training, observation of experts, routines, meetings, standard operating procedures, manuals and databases where most of transferring knowledge process is the implication of strategic alliances, joint ventures, mergers and acquisitions. KT especially through strategic alliances has become a shot gun approach for a firm to acquire knowledge that it could not easily develop within its confines. One of the strategic alliances practices in Malaysia is through IT outsourcing. Public agencies can increase knowledge and diversity through outsourcing. However, this will not necessarily translate into increased organizational knowledge if the organizations failed to assign value to the knowledge they are transferring and receiving from the partnership.

During IT outsourcing partnership, client and vendor can develop two forms of knowledge transfer in terms of a reciprocal learning [13]; 1) the partners can obtain from each other technical knowledge and know-how, 2) they can learn from each other management and business skills that individually they are lacking. Both the service receiver and provider should have a shared vision and goals for partnership as well as a belief that their partners will not act opportunistically [14]. Knowledge transferring or sharing throughout the IT outsourcing progress management should be given more attention for both sides. One side, vendors can transfer their IT special knowledge to clients, which helps client to improve their IT function; on the other aspect, clients also transfer their business knowledge to vendors, which will improve vendor's capability of understanding and implementing. Unfortunately, it appears that public sector organizations in developing countries especially Malaysia, have not received much attention in the research literature covering knowledge transfer especially in IT outsourcing. Most of the studies concentrate on the general knowledge management implementation or readiness at public agencies [15], Malaysian SME industries [16], aerospace industry [17], bank [18], telecommunication industry [19], higher education [20] to cite a few. There is only one work recently done by [21] focusing on knowledge transfer success factors in Malaysia setting. From the success factors the authors developed a theoretical framework for future work. Apparently, those researches never address the need of organizational learning context for an effective knowledge transfer. Therefore, it is crucial for this study to be taken and significantly give an insight and better understanding of the knowledge transfer processes in ITO.

2.2 Malaysia E-Government IT Outsourcing

In today's world, governments are increasingly under pressure for more profound change in structure and strategies to meet the requirements of contemporary society. Government needs to become more partnership-based, results-oriented, integrated, and externally focused. Therefore, government starts to serve their citizen thru electronic application. Malaysian government has starts their initiative in transforming their service delivery by launching seven flagship of e-government with the development of Multimedia Super Corridor (MSC) has become a jump starts of all current transformation. In order to focus more on servicing citizen, e-government outsourcing has become an important measure to reduce the pressures from cost, technical, as well as personnel. E-government Outsourcing in the Malaysia public sector has become an accepted management practice. Yang et al. [22] classified e-government outsourcing into two types; (i) system construction outsourcing (project in nature) and (ii) maintenance outsourcing (process in nature). Usually e-government outsourcing project will involve two or more vendors working together for one particular project. The relatively high complexity, high uncertainty, and high risk of large e-government service projects favour a partnership approach. This government (clients)-private (vendors) partnership make the knowledge transferring process more problematic due to differences in the development and implementation of IS across sectors.

According to a joint publication by Outsourcing Malaysia and ValueNotes published in August 2009, revenues from the Malaysian ITO industry are expected to touch $1.1 billion in 2009. The industry is expected to grow at a CAGR of 15% to reach $1.9 billion by 2013. Currently, ITO services in Malaysia have a greater share of the overall outsourcing market, followed by Business Process Outsourcing (BPO) services; while knowledge services outsourcing is still in its nascent stage, has a smaller share. The interest in outsourcing is still growing especially among players in the banking (e.g: CIMB & Maybank), airline (Malaysia Airline System), manufacturing, healthcare, and government sectors. IT outsourcing has been identified as one of the main ways to address some demanding challenges faced by government. The shortage of IT expert and the difficulty of attracting and retaining the right IT talent ranked as the number one barrier that fuel the Malaysian government decision to outsource. Current e-government IT outsourcing activities in Malaysia are data entry, ICT hardware maintenance, network management service, web-hosting management and development and application system maintenance [2]. However, there is a trend for government and public agencies to shift to more interactive service delivery which are citizen-centered and based on networks and partnership between public, private and NGO and between levels of government. The use of application providers by government can help meet increasing e-government service demands by citizen and business alike.

Currently, Malaysian government has been practicing three types of IT outsourcing model for e-government application namely [2]; (1) BOT (*Build, Operate, Transfer*), (2) BOO (*Build, Operate, Own*) and (3) Contract Services. For *BOT* approach the provider or vendor need to develop the application according to the agencies requirement and manage the system operation for a certain time as stated in the contract. After the contract terminate, the vendor will hand over the application to the agencies that owned the project. Example applications for BOT approach that have been implemented are e-procurement (e-perolehan) own by Ministry of Finance (MOF) and The Electronic Budget Planning and Control System (e-SPKB) own by National Accountant Department (ANM). In contrast with *BOO* outsourcing approach, the vendor will provide and manage the ICT service without hand in back to the agencies. The ownership of the services is still under vendor supervision. The last outsourcing approach is *contract basis service*. For this approach, the owner agency will give a contract to the vendor to develop or maintain the whole ICT devices but the ownership of the device belongs to the agencies not the provider. Most of Malaysia Public Agencies ITO contracts were three years or less since this duration had higher success rate compared to contract duration greater than three years [23].

Malaysian government has massively outsourced many e-government applications but scarce researches have focused on knowledge transfer processes in the outsourcing projects particularly for Malaysia environment. Although most of the success factors for ITO were rigorously considered based on principles and findings from previous research, which are frequently referred to [24], there are still some project that is not fully satisfied by the stakeholders or do not meet stated performance objectives [25]. Report from egov4dev.org (2009) has shown that e-government project failed because there is no lesson learned since knowledge about the failure was not captured, transferred or applied. As a result, mistakes were wastefully repeated. This claimed was also supported by [26] which examined the importance of knowledge transfer towards vendor's development that can create added value to the organizations. Giannakis [26] asserts that the failure of many initiatives revealed a twofold problem: first there is great difficulty in the generation and transformation of knowledge into organizational action and subsequently and even greater difficulty in the transfer of knowledge to partners. In addition, the acquired application may not be customized enough to effectively streamline or transform the business process. Moreover, this relates to the criticism that the vendors have limited understanding of the clients' business process [27]. IT outsourcing involves integrating and coordinating knowledge from many individuals of different disciplines and backgrounds, with varied experiences and expectations, located in different parts of the organization. Thus, both client and vendor should able to identify types of knowledge that is needed to be transferred during project execution, what mechanisms are appropriate and how the transferred knowledge can be retain in the organization for learning purposes and future use. To address the issues, we have drawn our research from two popular theories in Knowledge Management field as well as outsourcing field.

3. Literature Review

3.1 Theoretical Lens

According to [28], the popular theories being used in ITO research is the economic theory (e.g. Transaction Cost Theory & Agency Theory), followed by sociology theory (e.g. Relational Exchange Theory & Social Exchange Theory) and lastly strategic management theory (e.g. Resource-Based Theory, Resource Dependence Theory). From the researcher literature review, for the past five years research in ITO and knowledge transfer, most researchers used multiple theoretical approaches rather than single theoretical approach. The most dominated theory behind the knowledge management activities in ITO project were two popular models; Resource-Based View Theory (RBV) and Knowledge-Based View Theory (KBV). From a sourcing perspective, RBV theorists have traditionally maintained that firms should not outsource any business function or activity that contributes to building and maintaining competitive advantage. According to this two theories postulated by [29] and [30], firms that established connections with external firms through mechanisms such as outsourcing run the risk of transferring vital knowledge and resources by engaging in sourcing partnerships. Other potential negative sourcing outcomes include creating competitors via vertical integration of sourcing partners and losing vital internal knowledge and resources by engaging in sourcing relationships with external partners. As a result, RBV called for a protectionist stance regarding outsourcing, recommending that firms should only outsource support functions that do not directly contribute to the firm's value added and competitive advantage generating mechanisms.

From a more proactive perspective, RBV and KBV tenets denote that firms may engage in outsourcing as a means of identifying, exploring, and transferring knowledge and resources from external sourcing partners to internal control. KBV proposes that IT outsourcing is a way to utilize vendor's professional knowledge and skills [31]. Although the knowledge-based view emphasizes the unique knowledge of the client firm [32], IT projects needs an integration of mix experience and new knowledge from the vendor. Client and vendor firms can create shared understanding from a successful exploration of specialized external knowledge. The exploration of external knowledge in IT outsourcing needs a knowledge integration of client domain knowledge and vendor technical knowledge during the development process. Without such integration, the unique knowledge of the client firm cannot be successfully leveraged in the outsourced custom-software development process. Consequently, IT outsourcing can be viewed as a boundary crossing mechanism through which firms can use sourcing relationships to gain access to resources critical to the firm's competitive advantage development or maintenance [33]. In such cases, client establishes a short-term relationship with an established outsourcing partner with the intent of transferring knowledge, human capital, and technologies from the client to the vendor. Additionally, [33] asserts, mechanisms emphasized in outsourcing strategy can range from the (i) transferring of knowledge to help develop internal capabilities, (ii) by the hiring an experts personnel from the sourcing firm to build up internal capabilities for the partner, (iii) by the outright acquisition of the sourcing firm to internalize capabilities previously existing externally and lastly (iv) by aligning client's needs with vendor possessing complementary resources and capabilities. In such cases, outsourcing partners may provide the combination of complementary knowledge bases with a lack of direct competition that can fuel innovation of a new application/technology and service development. Hence, RBV and KBV perspectives provide valuable insights for the business rationale of IT outsourcing practices. Many researchers have found them useful in explaining specific aspects of outsourcing decisions, processes and outcomes using KBV and RBV theoretical lens. Thus, many researchers [39],[35] have placed these two theories as the theoretical lens to the KT model or their framework specifically for ITO environment.

3.2 Knowledge Transfer Model

King et al. [4] appointed two important element in developing effective organizational knowledge; (i) communication and (ii) information processing. There are three models dominated within the knowledge transfer area. Most of the existing KT models were rooted from communication model, group information processing model and knowledge creation model. Communication based model was elucidated by [36] and later being improvised by [37] while the second is based from Hinsz's [38] model. The third one is based from Nonaka's [39] knowledge creation model. Within the communication-based approach, the transfer of knowledge is regarded as a message encoded in a medium by a sender to a recipient in a given context. Schramm's [36] communication model initially consisted of three elements; (i) Sender, (ii) Recipient and (iii) Message. The receiver becomes the "recipient" or "user", since it is the subject who learns or acquires knowledge (not simply the message receiver) whereas; the "sender" is the knowledge holder. The message becomes the "object", as it can be produced by complex knowledge. Scharmm's [40] later enhanced the model by including Media. Media is the channels used to communicate the message, palliate its passage, and enhance its chances of completing a communicative act. Scharmm's [40] model becomes the most referred basic model in many knowledge transfer framework. Subsequently, [37] improvised the basic model developed by Schramm's [40] by considers six factors: Knowledge source, Message, Knowledge receiver, Channel, Feedback and Environment or Organizational context.

Apart from viewing KT from communication lens, scholars started to integrate the communication model with group information processing model to enhance the existing KT model. In order for the organization to learn something, the members need to process the data or information that they got to better suit the organization. Hinsz et al. [38] has postulated three components in the information processing model: *encoding* (i.e. forming knowledge representations through interpretation, evaluation and transformation), *storing* (i.e. entering representations in the memory system), and *retrieval* (i.e. accessing and using representations from the memory system). This concept is closed to human cognitive system. Later, Gibson [41] starts to improvised Hinsz's model by expending the information processing into four stages; accumulation, interaction, examination and accommodation. However, Gibson's model is applicable if the accumulated knowledge is highly ambiguous and the processing does not occur in a linear time order. The main similarity between these two models is the need of social interaction along each phase.

From [42] framework of knowledge generation, the transfer of knowledge is seen as the creation of knowledge through four modes of knowledge conversion of explicit and implicit forms of knowledge: externalization (from implicit to explicit), combination (from explicit to explicit), socialization (from implicit to implicit) and internalization (from explicit to implicit). Nonaka and Takeuchi [42] visualized the knowledge conversion process as cyclic process and happen mainly through informal networks of relations in the organization starting from the individual level, then moves up to the group (collective) level and eventually to the organizational level. However, according to [43], Nonaka's model does not describe how to initiate the macro level process for individuals or groups to manage the knowledge and to be innovative. Gourly [44] further claimed that Nonaka only proposed two modes of knowledge creation; internalization and externalization, whereas; socialization and combination are modes of knowledge transfer. Based from the discussed constraint, [45] develops an integrative cognitive architecture model for groups with the combination of three subsystem; selection subsystem, memory subsystem and communication subsystem. Curseu [45] claimed that the comprehensive group information processing models should consist of communication based view, knowledge creation based and memory based system. The proposed information processing model can be integrated with communication model, Gibson's model, Nonaka's Model and Transactive Memory theory. This model is suitable at the organizational level unit of analysis for example this model appropriate for distributed group members and virtual project team. However, the group members must be anonymous.

Besides the three basic models as the basis to the KT model developed by past researchers, scholars have also embodied KT antecedents and consequences in the model. Prior studies have investigated the role of knowledge characteristics, such as ambiguity, in determining knowledge transfer [10]. Other studies have examined sender-receiver characteristics, such as absorptive capacity and motivation [46],[47] or organization context [48],[49]. Inspite of that, current trends in knowledge transfer research have also comprised project nature [5][35] factors in developing the model since most of the transferring process occurred during the project execution or alliances. Table 1 summarized a few KT components that being derived from the past research. These components have been reviewed by most of the scholars in KT research and significantly gives effect on KT process in ITO.

In spite of all factors discussed above, organization information system is claimed to be an effective tools to support knowledge transfer process. However, most of the organizational knowledge is based on the information stored in legacy information systems which have been developed in an isolated way [59]. Therefore, such information can be inconsistent, redundant and difficult to retrieve and link. The information that ends up in the most organizational information system has a poor structure (e.g., PDF documents), which makes the system unmanageable and chaotic, limiting the possibility to deal with other system requirements, such as information privacy and fast and flexible retrieval methods [59]. It was suggested that the organizational information memory system should have the capability to provide an experts database with points of contact on various topics [60], support both formal and informal knowledge besides the automatic privacy mechanism [59]. Hence, recent scholars have connected organizational information memory system (OMIS) with the Transactive Memory System (TMS) to facilitate the interaction of organizational knowledge [61],[62],[63].

Table 1: Knowledge Transfer Component

Components	Characteristics	Authors
Source	Disseminative Capacity Reliability Credibility Willingness to share	[10],[35],[50],[47],[49]
Recipient	Absorptive Capacity Motivation Learning intent Retentive Capacity	[50],[51]
Knowledge	Knowledge Ambiguity Stickiness Complexity Tacitness	[47],[51],[52]
Organizational	Organizational Culture Personnel Movement Community of Practices Management Practices Organizational Structure Organizational Learning Strategy	[53],[54],[55] [56]
Communication	Codification Interpretation Communication Channel	[57]
Relationship	Arduous Relationship Dyadic relation Strength of ties Network density Social Similarity	[10],[58],[57]
Project Nature	Prior collaboration history Team size Project complexity Project phase	[35]

3.3 Organizational Memory System and Knowledge Transfer: The Role of Transactive Memory System (TMS)

KT process comprises four activities; knowledge conversion, knowledge routing, knowledge dissemination and knowledge application [64]. Within these practices, effective transfer and use of organizational knowledge depends to a large extent on the organization's ability to create and manage its collective memory. The organization itself has been seen as a repository of knowledge [65]. The organization's knowledge repositories or knowledge stock are found in individual members, roles and organizational structures, standard operating procedures and practices, culture and physical layout of the workplace [64]. This collective memory is often referred to as organizational memory (OM). To support effective management of organizational memory, [66] proposed the use of information technology to accomplish four specific processes related to organizational memory: acquisition, retention, maintenance, and search and retrieval. In addition, they outline a design for an organizational memory information system (OMIS).

However, [67] argued that the proposed OMIS architecture by [66] faced several challenges. According to them, much of the knowledge in the OM is contextualized and consequently the knowledge interpreted wrongly. A second challenge regarding the locations of knowledge since OM generally resides in different types of retainers. These retainers of OM may be in dispersed location and their memories might be difficult to combine. A third problem with OM management is that knowledge is often tacit which is difficult to track and maintain in large organizational memories. A fourth problem concerns with the unpredictability of organizational knowledge. This unpredictability results in frequent changes to the contents of the OM measure of the retainer’s legitimacy and reliability is required. These five problems create difficulties for members of the organization in retrieving and using knowledge that resides in OM. Therefore, to gain a better understanding of possible ways to overcome the barriers for efficient OM management, [67] proposed the concept of Transactive Memory Systems (TMS) being incorporated with OM.

One of the philosophical theories that have been embedded in the concept of organizational memory is Transactive Memory theory. Transactive Memory theory becomes Transactive Memory System when [68] started to model human memories in a concept of computer network. Transactive memory is a system for encoding, storing, and retrieving information in groups [68]: it is a set of individual memory systems in combination with the communications that take place between individuals. Originally, TMS was used to describe the ways in which dyads (such as married couples) that are close to one another share knowledge and allocate responsibilities for knowing. Extending the notion of TMS beyond groups and pairs, several authors have speculated on how organizations might function as TMS with an input of information system architecture. Anand et al. [46] proposed certain forms of information systems, such as intranets, search engines, standardized concepts and vocabularies, could be used to enhance the functioning of TMS. Nevo and Wand [67] proposed directories of meta-knowledge to overcome the knowledge storage and location problems as stated before. The computerized directories of meta-memory can compensate for the lack of the group’s tacit knowledge. Even so, the work on organizational TMS has been conceptual rather than empirical. There have been no descriptions of working organizational TMS in the literature.

Therefore [62] have proposed a model of the operation of an organizational TMS. This model focused more on organizational KM codification strategy rather than personalization strategy since the aim of suggested model was to connect people with reusable codified knowledge. Jacobson and Klobas [62] have divided organizational TMS into four main activities instead of three activities postulated by [68]. The nucleus of organizational TMS is the directory or the knowledge repositories. The directory consists of metadata about people, including name, organizational role and formal group membership, work experience, areas of expertise and other information such as availability and reliability as a source of knowledge. Some of the metadata for some people in a TMS will be stored in a person's head, but other metadata can be stored externally, in a CV or expertise database, a document management or knowledge management system, on the organization's intranet or in handbooks, or in the heads of intermediaries such as managers, administrators and other colleagues who act as gatekeepers or links in a chain to the ultimate source of the knowledge. The second activity is directory maintenance. According to them, directory can be maintained by formal and informal procedures. Formal procedures might include the updating of metadata and other information in organizational information systems whereas informal procedures include discussions held alongside formal meetings or serendipitous meetings in the corridor or coffee room. The third activity is retrieving process from the directory. The directory allows knowledge to be retrieved from one's own work group(s) and from others in the organization. Much of the information retrieval from one's own group might be in the form of conversations although this retrieval might be supported by information systems that record knowledge in the form of documents. Finally, knowledge allocation would be the fourth activity evoked [62]. They argued that knowledge is allocated and stored on the basis of several activities ranging from formal allocation of responsibility and transfer of knowledge among people in the organization to individual learning. This view provides a framework to guide development of a holistic TMS for a particular organization. It allows a view of what an information system might provide and what is best done (or indeed must be done) through interpersonal means.

4. Conceptual Framework

The underpinned framework for this study is derived from the in-depth study on IT/IS outsourcing, knowledge transfer, information processing literature and organizational learning. Previous research has examined a range of antecedents of organizational knowledge transfer. For this research purposes, this study included only antecedents that have been studied extensively across multiple studies and align conceptually. This enabled researcher not only to compare antecedents, but also to make sure the antecedents studied are deemed relevant by the research community. Consistent with prior literatures, the researcher classifies antecedents of inter-organizational knowledge transfer into four domains: organization memory system factors, client-related factors, vendor-related factors while project management factors as controlled variables. This paper contributes to the existing literature by examining how organization memory system can facilitate the knowledge transfer process between client and vendor involved in IT outsourcing relationship besides the other three most cited determinants. From the IT project management perspectives, organization shared cognition are able to successfully manage project interdependencies [69]. Fig. 1 illustrates the proposed conceptual framework for the study.

4.1 Variables

The dependent variable in the research framework is ‘knowledge transfer’. The operationalize definition of knowledge transfer for this research was drawn upon the communication theories, whereby transfer of knowledge is define as a method that involves two-way communication between the client and the vendor exchange and share their useful information/ skill/ competencies or routines about the project and both parties is affected by changes in recipient replication and adaptation capabilities and changes in skills/knowledge. Knowledge from this research context is organizational knowledge whereby “knowledgeable” organization can be seen through daily basis routines and the systematic structure of workflow. A vendor corresponds to the knowledge source involved in transferring knowledge or the generalized knowledge resource, whereas; client act as knowledge receiver, and the destination or the entity which receives and internalizes the knowledge content. Further, within the knowledge transfer context, the transmission element corresponds to the activities and processes, such as communication activities, through which knowledge is transferred from one entity to the other.

Meanwhile, the independent variables are measured by three domains; vendor characteristics, client characteristics and organizational memory context. Each of the domains is observed by several items that have been selected from Table 1. Researchers only take the items that empirically give significant or positive impact towards knowledge transfer. The negative impact has been eliminated to ensure the high validity and reliability of each construct. Client in this research context is the Malaysian public agencies act as the recipient of knowledge that outsourced the E-government application to the third parties. Meanwhile, vendor is conceptualize as a third-party entity act as the source of knowledge that develop, manages and distributes E-government application and solutions to public agencies. Vendor characteristics are measured by vendor credibility, willingness to share, disseminative capacity and knowledge integration. For client characteristics, researchers have chosen four measurable item; absorptive capacity, retentive capacity, conjecture and motivation. Researchers have also incorporated Transactive Memory System (TMS) as proposed by [62] and [61]. Although the most popular measurement of TMS is elucidated by [70], by which TMS is measured by specialization, coordination and credibility, we argue that the early measurement developed by [71] is based from the activities memory recall of dyads that working together and it is most suitable of TMS form individual’s perspective rather from organizational perspective. Therefore, we have extracted the main organizational routines that involves during outsourcing project as presented in [61][71] and [72] interpretive research. Therefore, TMS is measured from the organizational project routines that encoding and updating directories, coordinating and retrieval, allocating and storing and lastly directory content.

Much of the academic research on information system project management has been done in the context of software development and maintenance in the “traditional” computing paradigm in which the majority of software projects involve the custom development of applications [35]. There is a lack of empirical investigation of the issues related to the IT outsourcing projects. Control variables in this model are derived from project management literature, but we labelled it as project nature. Thus in this research, four control variables are included in the framework: prior collaboration history, team size, project complexity and project phase.

4.2 Framework Hypotheses

Generally there are two main stream of KT approaches [73]; (1) Information Science approach (knowledge as an object) and (2) Constructivist approach (knowledge as process, a set of dynamic skills and know-how). The Western or Europe companies prefer Information Science approach. In contrary, Japan companies much emphasized on constructivist approach [74]. In this study, we developed our framework by integrating both approaches. We viewed TMS from Information Science approach. Within this approach, knowledge is viewed as an object that can be created, stored, and retrieved. Meanwhile, the other three variables (client characteristics, vendor characteristics and knowledge transfer) are from constructivist approach. In this approach, knowledge is primarily viewed as a process, a set of dynamic skills and know-how that is constantly changing. Constructivist approach is involved with assessing, changing, and improving IT outsourcing team skills and behaviour. We integrate both approaches depending on the type of knowledge in IT outsourcing project. IT outsourcing project consists of procedural and domain specific knowledge. Procedural knowledge is tacit in nature while domain specific knowledge is explicit in nature. The proposed conceptual framework is presented in Fig. 1. We derived 12 hypotheses from each constructs that relates.

4.2.1 Vendor Characteristics

Many studies have examined the effect of knowledge source on knowledge transfer. The knowledge source in this research refers to the vendor that develops or provides the e-government applications or infrastructures. There are four characteristics of vendor that being measured in this study; vendor's credibility, vendor's willingness to share, vendor's disseminative capacity and vendor's capability to integrate knowledge from various units/departments. Vendor's credibility is generally defined as the extent to which a client perceives a vendor to be trustworthy and reputable [35]. Thus from the definition, the credibility concept has two dimensions: trust and reputation. Knowledge transfer researchers have indicated trust as the core ingredient in order for individuals to transfer knowledge [10]. Trust 'reflects the belief that a partner's word or promise is reliable and that a partner will fulfil its obligations in the relationship' [65]. When client credibility is high, client are likely to be more open and receptive to information from the vendor; ideas in the asset are perceived to be worthy of consideration. The knowledge conveyed is thus more likely to be seen as useful, and to influence the behavior of the recipient [10]. The importance of a client's credibility is amplified in the context of a knowledge transfer process because this process is not amenable to enforcement by contract [75].

Besides vendor's credibility, we also measures knowledge sharing initiatives in the project. Lee [14] and [35] have showed that knowledge sharing is a major indicator of whether or not the outsourcing activity succeeds. Those studies confirms that knowledge sharing is one of the major predictors for outsourcing success because IT outsourcing posses highly valuable knowledge relating to the product development, the software development process, project management and technology in general [76]. Therefore we operationalized willingness to share as vendor attitude which vendor is willingly to provide access towards others about knowledge and his experiences. Willingness to share is operationalized based on the intensity level of vendor in doing tacit and explicit knowledge sharing with his client in ITO project. Willingness to share also relates to the vendor's disseminative capacity.

Disseminative capacity refers to the vendor capacity to contextualize, format, adapt, translate and diffuse knowledge through a social or technological network and to build commitment from stakeholders [77]. In the context of IT outsourcing, individual members who control and distribute resources, information and knowledge can largely affect the performance of the whole project team [78]. The fourth constructs is knowledge integration. Knowledge integration is defined as individual members who control and distribute resources, information and knowledge can largely affect the performance of the whole project team [78]. In an IT outsourcing project, the users from the client organization communicate system requirements to the vendor's IT consultants who use their software expertise and knowledge from the users to build the system. Users then assimilate the system by making necessary changes to their work. Knowledge integration is essential since if knowledge from a particular cluster is missing or is not integrated. Therefore, we derived four hypotheses from vendor characteristics:

H1a: Vendor credibility significantly gives an impact towards vendor characteristics for knowledge transfer processes in IT outsourcing at public agencies

H1b: Vendor willingness to share significantly gives an impact towards vendor characteristics for knowledge transfer processes in IT outsourcing in public agencies

H1c: Vendor disseminative capabilities significantly give an impact towards vendor characteristics for knowledge transfer processes in IT outsourcing in public agencies

H1d: Vendor knowledge integration significantly gives an impact towards vendor characteristics for knowledge transfer processes in IT outsourcing in public agencies

4.2.2 Client Characteristics

Second independent variables involved in this research are the client factors. There are four independent variables has been identified in this research that influence the process of transferring knowledge in IT outsourcing project; absorptive capacity, retentive capacity, communication competence and motivation. Most scholars stress that the studies of knowledge transfer should concern not only whether knowledge owners have a willingness to share, but also whether knowledge receivers can learn and absorb. Therefore, absorptive capacity affects vendor ability to recognize the importance and value of new knowledge, to assimilate the knowledge, and to apply it to solve the problem. We defined absorptive capacity as the ability of the client to acquire new external knowledge, assimilate or transform the knowledge into usable knowledge then apply it to business ends. The definition emerges two subsets; potential absorptive capacity and assimilation and realized absorptive capacity. The client needs to actually know their prior knowledge and their ability to valued new knowledge that they received from vendor for example through training or project maintenance. A transfer of knowledge is effective only when the knowledge transferred is retained. The ability of a recipient to institutionalize the utilization of new knowledge reflects its 'retentive' capacity [10]. According to [7], clients retentive capacity is differs from absorptive capacity because absorptive capacity refers to an indication of initial short-term memory, whilst; retentive capacity refers to long-term memory.

Communication competence can be defined as the extent by which the client and vendor have a frequent routine of formal (in term of task-achieving issues) or informal (out of role) interaction and conversation regarding project-relevant information. The uncertainty situation in IT outsourcing, my impacts the process of knowledge transfer among clients' and vendors' that emerged the important of communication competence. On top of that, client needs the motivation to accept and absorb the new external knowledge. The motivation of the client refers to the client desire to implement the knowledge or technology being transferred. Lack of motivation in knowledge transfer will result in passiveness, feigned acceptance or implementation, hidden sabotage, intentionally slow implementation, or directly reject the practice. From the above argument, we posit four hypotheses from client variables;

H2a: Client absorptive capacity significantly gives an impact towards client characteristics for knowledge transfer processes in IT outsourcing in public agencies

H2b: Client retentive capacity significantly gives an impact towards client characteristics for knowledge transfer processes in IT outsourcing in public agencies

H2c: Client communication competence significantly gives an impact towards client characteristics for knowledge transfer processes in IT outsourcing in public agencies

H2d: Client motivation significantly gives an impact towards client characteristics for knowledge transfer processes knowledge transfer in IT outsourcing in public agencies

4.2.2 Organizational Memory Context: TMS

Knowledge transfer in group encompasses various practices of managing organizational knowledge. Effective transferring and use of organizational knowledge depend on a large extent of the organization's ability to create and manage its collective memory. This collective memory is often referred to as organizational memory (OM). In relation to IT outsourcing, such memory resides in business professionals from clients' side and IT specialist from vendors' side, policies, contract/agreement, and culture. These retainers of OM may be in different locations and their memories might be difficult to combine. There are three processes of TMS that affects the knowledge transfer within IT outsourcing team [79]; first, the directory updating functions allows group members to be aware of the location of the expertise possessed by specific individual. Secondly, information allocation and function represents the process of distributing knowledge to the members whose expertise is best suited for its storage. Third the retrieval coordination function shows how to retrieve needed information on any topics based on related knowledge from individual expertise in the memory system. In this paper, we enhanced the TMS concepts by incorporating other information processing activities like encoding, coordinating, storing and relevant organizational directories content to support the knowledge transferring processes. Thus, scholars have increasingly considered the concept of the TMS as an enhancer of inter-organizational knowledge transfer [61][67] and to develop organizational knowledge memory system [62]. While the concept of TMS has been studied in the context of traditional organizational forms and co-located teams, little is known about the process through which a TMS in distributed teams could be created and could support knowledge transfer between remote sites like a case in IT outsourcing. We measures TMS in terms of the project routines to encoding and updating the directories, coordinating and retrieval process, allocating and storing of project information or data in the organizational memory systems. Thus, we believe that TMS will facilitate the knowledge transferring process;

H3a: The relevant organizational directories content will facilitate knowledge transfer in IT outsourcing at public agencies

H3b: The project routines of encoding and updating project document will facilitate knowledge transfer in IT outsourcing at public agencies

H3c: The project routines of coordinating and retrieval project document will facilitate knowledge transfer in IT outsourcing at public agencies

H3d: The project routines of allocating and storing project document will facilitate knowledge transfer in IT outsourcing at public agencies

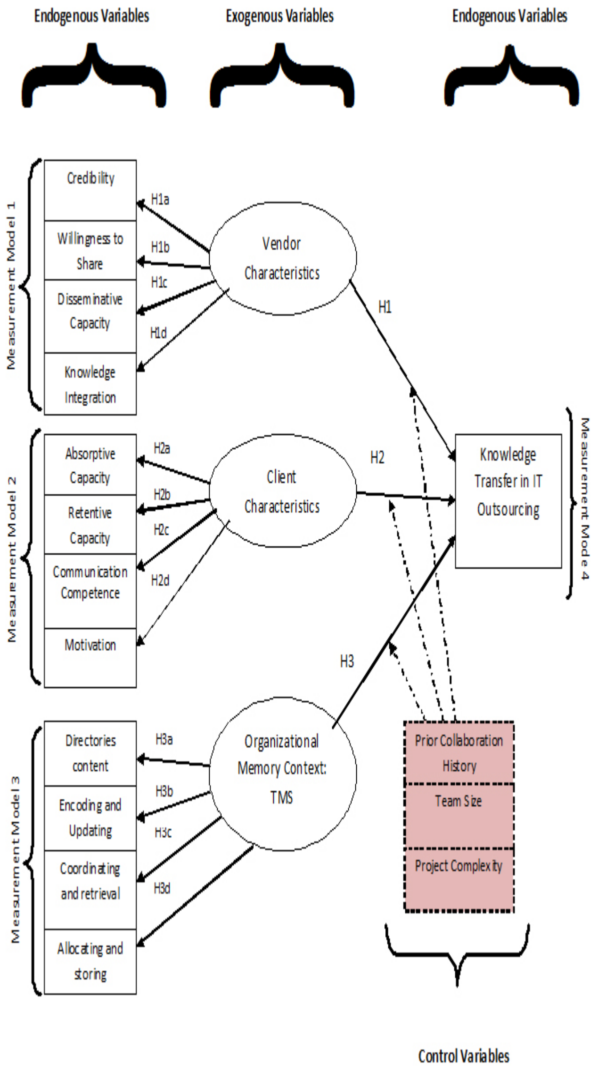


Fig. 1 Proposed Conceptual Framework.

5. Final Remarks

This conceptual paper proposed an integrative preliminary framework that links four groups of key domains namely; client related characteristics, vendor related characteristics, Transactive Memory System context and project nature factors while discussing the theories and models behind the proposed model. This conceptual model is still based on literature study. Therefore, it needs further research to empirically validate the model. We believed that the application of the framework may provide useful insights into ITO specifically for Malaysia e-government initiatives.

References

[1] B. Hacket, “Beyond Knowledge Management: New Ways to work and Lear”, The Conference Board, Research Report, 2000, 1262-00-RR.

[2] MAMPU. *Malaysia Public Agencies IT Outsourcing Guideline*. Retrived November 20, 2009, from http://www.mampu.gov.my/pdf/Garis-Panduan-IT-outsource.pdf

[3] Z. Yusof, and M. B. Ismail, “Is There A Relationship Between Knowledge Sharing Practice And The Quality of Service Delivery? A Case Study In Three Governemnet Agencies In Malaysia”, Journal of Konwledge Managemeny, Vol. 10, No. 1, 2009.

[4] R. C. King, W. Xia, J. C. Quick, and V. Sethi, “Socialization and organizational outcomes of information technology professionals”, Journal of Career Development International, Vol. 10, No. 1, 2005, pp. 26-51.

[5] D. G. Ko, L. J. Kirsch, and W. R. King “Antecedents of knowledge transfer from consultants to clients in enterprise system implementations”, MIS Quarterly, Vol. 29, No. 1 , 2005, pp. 59-86.

[6] L. Argote, B. McEvily, and R. Reagans, “Introduction to the Special Issue on Managing Knowledge in Organizations: Creating, Retaining, and Transferring Knowledge”, Management Science, Vol. 49, No. 4, 2003, pp. 5-8.

[7] D. G. Schwartz, “Integrating knowledge transfer and computer-mediated communication: categorizing barriers and possible responses” Knowledge Management Research & Practice, Vol. 5, Issue August, 2007, pp. 249-259.

[8] J. A. Kumar, and L. S. Ganesh, “Research on knowledge transfer in organizations: a morphology”, Journal of Knowledge Management, Vol. 13, No. 4, 2009, pp.161-174.

[9] T.C. Ambos, and B. Ambos, “The impact of distance on knowledge transfer effectiveness in multinational corporations”, Journal of International Management, Vol. 15, No. 1, 2009, pp.1-14.

[10] G. Szulanski, “The process of knowledge transfer: a diachronic analysis of stickiness, Organizational “, Behaviour and Human Decision Processes, Vol.82, No. 1, 2000, pp. 9-27.

[11] D. C. Mowery, J. E. Oxley, and B.S. Silverman, “Strategic Alliances and Interfirm Knowledge Transfer”, Knowledge Creation Diffusion Utilization, Vol. 17, 1996, pp. 77-91.

[12] A. M. Ghassani, “Improving the Structural Design Process: a Knowledge Management Approach”, Unpublished doctoral dissertation. Loughborough University, 2003.

[13] J. Connell, and R. Voola, “Strategic alliances & knowledge sharing: Synergies or silos?” Journal of Knowledge Management , Vol. 11, No. 3, 2007, pp. 52-66.

[14] J. N. Lee, “The impact of knowledge sharing, organizational capability and partnership quality on IS outsourcing success”, Information and Management, Vol. 38, No. 5, 2001, pp. 323-335.

[15] S.O.S Syed-Ikhsan, and F. Rowland, “Benchmarking knowledge management in a public organisation in Malaysia”, Benchmarking: An International Journal, Vol. 11, No.3, 2004, pp.238-266.

[16] K. W. Wong, "An exploratory study on knowledge management adoption in the Malaysian industry", International Journal of Business Information System, Vol. 3, No. 3, 2008, pp. 272-283.

[17] L.W. Tat, and S. Hse, "Knowledge Management in The Malaysian Aerospace Industry", Journal of Knowledge Management, Vol.1, No. 1, 2007, pp. 143-151.

[18] H. M. Ali, and N. H. Ahmad, "Knowledge Management in Malaysian Banks: A New Paradigm", Journal of Knowledge Management Practice, Vol. 7, No. 3, 2006.

[19] C.C. Wei, C.S. Choy, and W.K. Yew, "Is the Malaysian telecommunication industry ready for knowledge management implementation? ", Journal of Knowledge Management, Vol. 13, No. 1, 2009, pp. 69 – 87.

[20] S.D. Ramachandran, S.C. Chong, and H. Ismail, "The practice of knowledge management processes: A comparative study of public and private higher education institutions in Malaysia", Vine, Vol. 39, No. 3, 2009, pp. 203-222.

[21] A. Mohamed, N. H. Arshad and N. A. Abdullah, "Influencing factors of knowledge transfer in IT outsourcing", Proceedings of the 10th WSEAS international conference on Mathematics and computers in business and economics, 2009, pp.165-170.

[22] J. Gao, W. Yang, and G. Fang, "A study on government-dominated e-government outsourcing model", Proceedings of the 3rd International Conference on Theory and Practice of Electronic Governance - ICEGOV '09, 2009, 235. New York, New York, USA: ACM Press.

[23] M. C. Lacity, S. A. Khan, and L. P. Willcocks, "A review of the IT outsourcing literature: Insights for practice", Journal of Strategic Information Systems, Vol. 18, No. 3, 2009, pp.130-146.

[24] J. Moon, G. Jung, M. Chung, and Y. C. Choe, "IT outsourcing for E-government: Lessons from IT outsourcing projects initiated by agricultural organizations of the Korean government", 40th Annual Hawaii International Conference on System Sciences (HICSS'07), 2007, pp. 104a.

[25] R. T. Nakatsu and C. L. Iacovou, "A comparative study of important risk factors involved in offshore and domestic outsourcing of software development projects: A two-panel Delphi study", Information and Management, Vol. 46, No. 1, 2009, pp. 57-68.

[26] M. Giannakis, "Facilitating learning and knowledge transfer through supplier development", Supply Chain Management: An International Journal, Vol. 13, No. 1, 2004, pp.62-72.

[27] Y. Chen, and J. Gant, "Transforming local e-government services: the use of application service providers", Government Information Quarterly, Vo. 18, No.4, 2001, pp. 343-353.

[28] T. Benedikt, M. Frank, "Why risk management matters in it outsourcing – a systematic literature review and elements of a research agenda", 17th European Conference on Information Systems, 2009, pp. 1-13.

[29] J. Barney, "Firm Resources and Sustained Competitive Advantage", Journal of Management, Vol. 17, 1991, pp. 99-120.

[30] B. Wernerfelt, "The resource-based view of the firm: Ten years after", Strategic Management Journal, Vol. 6, No.3 1995, pp. 171-174.

[31] M. Li and D. Li, "A Survey and Analysis of the Literature on Information Systems Outsourcing," Pacific Asia Conference on Information Systems, 2009.

[32] R. M. Grant, "Toward A Knowledge-Based Theory of The Firm", Strategic Management Journal, Vol. 17, Winter Special Issue, 1996, pp. 109-122.

[33] J. Combs and T. Crook, "Sources and consequences of bargaining power in supply chains", Journal of Operations Management, Vol. 25, No. 2, 2007, pp. 546-55.

[34] S. Blumenberg, H. Wagner and D. Beimborn, "Knowledge transfer processes in IT outsourcing relationships and their impact on shared knowledge and outsourcing performance", International Journal of Information Management, Vol. 29, No. 5, 2009, pp. 342-352.

[35] K.D. Joshi, S. Sarker and S. Sarker. "Knowledge transfer within information systems development teams: examining the role of knowledge source attributes", Decision Support Systems, 43, No. 2, 2007, pp. 322-335.

[36] W. Schramm, "The Process and Effect of Mass Communication", Urbana, University of Illinois Press, 1954.

[37] C.M. Jacobson. "Knowledge sharing between individuals, in Schwartz, D.G. (Eds), Encyclopedia of Knowledge Management", Idea Group Reference, Hershey, PA, 2006, pp. 507-14.

[38] V.B. Hinsz, R.S. Tindale and D.A. Vollrath, "The emerging conceptualization of groups as information processors", Psychological Bulletin, Vol. 121, No. 1, 1997, pp. 43-64.

[39] I. Nonaka, "A Dynamic Theory of Organizational Knowledge Creation", Organization science, Vol. 5, No. 1, 1994, pp. 14-37.

[40] W. Schramm, "Mass Communication: a Book of Readings", Urbana, University of Illinois Press, 1960.

[41] C.B. Gibson, "From knowledge accumulation to accommodation: cycles of collective cognition in work groups", Journal of Organizational Behavior, 22, 2001, pp. 121-34.

[42] I. Nonaka and H. Takeuchi, "The Knowledge-creating Company", Oxford University Press, New York, 1995.

[43] K. Sherif, and B. Xing, "Adaptive processes for knowledge creation in complex systems: The case of a global IT consulting firm", Information and Management, Vol. 43, No. 4, 2006, pp. 530-540.

[44] S. Gourlay, "Conceptualizing knowledge creation: a critique of Nonaka's theory", Journal of Management Studies, Vol. 43, No. 7, 2006, pp. 1415-1436.

[45] P. P. Curseu, R. Schalk, and I. Wessel, "How do virtual teams process information? A literature review and implications for management", Journal of Managerial Psychology, Vol. 23, No. 6, 2008, pp. 628-652.

[46] V. Anand, C.C. Manz and W.H. Glick, "An organizational memory approach to information management", Academy of Management Review, Vol. 23, No. 4, 1998, pp. 796-809.

[47] M. Easterby-smith, M.A. Lyles and E.W. Tsang. Inter-Organizational Knowledge Transfer: Current Themes and Future Prospects, Journal of Management Studies, Vol. 45, No. 4, 2008, pp. 677-690.

[48] U. Wilkesmann, H. Fischer and M. Wilkesmann, "Cultural characteristics of knowledge transfer". Journal of

Knowledge Management, Vol. 13, No. 6, 2009, pp. 464-477.

[49] R. Gregory, R. Beck and M. Prifling, “Breaching the Knowledge Transfer Blockade in IT Offshore Outsourcing Projects – A Case from the Financial Services Industry”, Proceedings of the 42nd Hawaii International Conference on System Sciences. IEEE, 2009, pp. 1-10.

[50] S. Sarker, S. Sarker, D. Nicholson, K.D. Joshi, “Knowledge Transfer in Virtual Systems Development Teams: An Exploratory Study of Four Key Enablers”, IEEE Transactions on Professional Communication, Vol. 48, No. 2, 2005, pp. 201-218.

[51] Q. Xu and Q. Ma, “Determinants of ERP Impllementation Knowledge Transfer”, Information and Management, Vol. 45, No. 8, 2008, pp. 528-539.

[52] L. Pérez-Nordtvedt, B.L. Kedia, D.K. Datta, and A.A. Rasheed, “Effectiveness and efficiency of cross-border knowledge transfer: an empirical examination”, Journal of Management Studies, Vol. 45, No. 4, 2008, pp. 714-744.

[53] L.Z. Cantu, J.R. Criado, and A.R. Criado, “Generation and transfer of knowledge in IT-related SME’s”, Journal of Knowledge Management, Vol. 13, No. 3, 2009, pp. 243-256.

[54] K.U. Ajmal, and M. M. Koskinen, “Knowledge Transfer in Project-Based Organizations: An Organizational Culture Perspective”, Project Management Journal, Vol. 39, No. 1, 2008, pp. 7-15.

[55] J. Rhodes, R. Hung, P. Lok, B.Y.-hui Lien, and C.-min Wu. “Factors influencing organizational knowledge transfer: implication for corporate performance”, Journal of Knowledge Management, Vo. 12, No. 3, pp. 84-100.

[56] C. Dhanaraj, M.A. Lyles, H.K. Steensma, and L. Tihanyi, “Managing tacit and explicit knowledge transfer in IJVs: the role of relational embeddedness and the impact on performance”, Journal of International Business Studies, Vol. 35, No. 5, 2004, pp. 428-442.

[57] B. Uzzi and R. Lancaster, “The role of relationship in inter-firm knowledge transfer and learning the case of corporate debt markets”, Management Science, Vol. 49, No. 4, 2003, 383-399.

[58] L. Agrote, P. Ingram, J. M. Levine and R. L. Moreland “Knowledge Transfer in Organizations: Learning from the Experience of Others”, Organizational Behavior and Human Decision Processes, Vol. 82, No. 1, 2000, pp. 1-8.

[59] S.F. Ochoa, V. Herskovic, and E. Pineda, “A transformational model for Organizational Memory Systems management with privacy concerns”, Information Sciences, Vol. 179, No. 15, 2009, pp. 2643-2655.

[60] C. H. T. Goh and V. Hooper, “Knowledge and information sharing in a closed information environment”. Journal of Knowledge Management, Vol. 13, No. 2, 2009, pp. 21 - 34.

[61] I. Oshri, P.C.V. Fenema, and J. Kotlarsky. “Knowledge Transfer in Globally Distributed Teams: The Role of Transactive Memory”, Information Systems Journal, 18, 2008, pp. 593-616.

[62] P. Jackson and J. Klobas. “Transactive memory systems in organizations: Implications for knowledge directories”, Decision Support Systems, Vol. 44, 2008, pp. 2409-424.

[63] J. Kotlarsky and M. Huysman, “Bridging Knowledge Boundaries in Cross- Functional Groups : The Role of a Transactive Memory System”, International Conference on Information System, 2009.

[64] B. Narteh, “Knowledge transfer in developed-developing country inter rm collaborations: a conceptual framework”. Journal of Knowledge Management, Vol. 12, No. 1, 2008, pp. 78-91.

[65] S. C. Currall and A. C. Inkpen, “A multilevel approach to trust in joint ventures”, Journal of International Business Studies, Vol. 33, No.3, 2002, pp. 479–495.

[66] E.W. Stein and V. Zwass, “Actualizing organizational memory with information systems”, Information Systems Research, Vol. 6, No. 2, 1995, pp. 85-117.

[67] D. Nevo and Y. Wand, “Organizational memory information systems: a transactive memory approach”, Decision Support Systems, Vol. 39, No. 4, 2005, pp. 549 – 562.

[68] D. M. Wegner, “A computer network model of human transactive memory”, Social Cognition, 13, 1995, pp. 1-21.

[69] M. Keith, H. Demirkan, and M. Goul, “Understanding Coordination in IT Project-Based Environments: An Examination of Team Cognition and Virtual Team Efficacy”, Proceedings of the 42nd Hawaii International Conference on System Sciences, 2009, pp. 1-8.

[70] K. Lewis, “Measuring transactive memory systems In the field: Scale development validation”, Journal of Applied Psychology, Vol. 88, No. 4, 2003, pp. 587-604.

[71] I. Oshri, J. Kotlarsky, and L. Willcoocks, “Global software development: Exploring socialization and face-to-face meetings in distributed strategic projects”, Journal of Strategic Information Systems, Vol. 16, No. 1, 2007, pp. 25-49.

[72] J. Kotlarsky and I. Oshri, “Social ties, knowledge sharing and successful collaboration in globally distributed system development projects”, European Journal of Information Systems, Vol. 14, 2005, pp. 37-48.

[73] J. T., Karlsen, L., Hagman, and T. Pedersen, “Intra-project transfer of knowledge in information systems development firms”, Journal of Systems and Information Technology, Vol. 13, No. 1, 2011, pp. 66-80.

[74] Juhana Salim, Nurul Rafidza Muhammad Rashid, Yazrina Yahya, Abdul Razak Hamdan, et al., “HiKMas: Cultural Behavioural and ontology based approach towards a Holistic Knowledge Management System Design”, Communication of The IBIMA, Vol. 8, 2009, pp. 107-113.

[75] J., Roberts, T., Analysis, and S. Management, “From know-how to show-how? Questioning the role of information and communication”, Technology Analysis & Strategic Management, Vol. 12, No. 4, 2000, pp. 429-443.

[76] A. Aurum, F. Daneshgar, and J. Ward. “Investigating Knowledge Management practices in software development organisations – An Australian experience”, Information and Software Technology, 50, No. 6, 2008, pp. 511-533.

[77] R. Parent, M. Roy, and D. St-jacques, “A systems-based dynamic knowledge transfer capacity model”, Journal of Knowledge Management, 11, No. 6, 2007, pp. 81-93.

[78] J. Mu, F. Tang, and D.L. MacLachlan, “Absorptive and disseminative capacity: Knowledge transfer in intra-organization networks”, Expert Systems with Applications, Vol. 37, No. 1, 2010, pp. 31-38.

[79] K.-ting, Zheng, T. C., Lin, and J. S. Hsu, "Understanding the Impact of Transactive Memory Systems on Project Team Performance : The Mediating Role of Knowledge Integration and Collective Mind", Information Systems, 2010, pp. 112-117.

Nor Aziati Abdul Hamid is PhD Candidate at Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia. She received her Master's Degree in 2006 at the Universiti Teknologi Mara Malaysia. She works as Lecturer at Production and Operation Management Department, Universiti Tun Hussein Onn Malaysia. Her research interests include knowledge transfer, Information System Outsourcing, E-Business and organization studies. Nor Aziati Abdul Hamid id the corresponding author for this paper.

Juhana Salim is a professor at Information Science program, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia. Her research interests include Knowledge Technology, Information Processing and Management.

Statistical Sign Language Machine Translation: from English written text to American Sign Language Gloss

Achraf Othman[1] and Mohamed Jemni[2]

Research Lab. UTIC, University of Tunis
5, Avenue Taha Hussein, B. P. : 56, Bab Menara, 1008 Tunis, Tunisia

Abstract

This works aims to design a statistical machine translation from English text to American Sign Language (ASL). The system is based on Moses tool with some modifications and the results are synthesized through a 3D avatar for interpretation. First, we translate the input text to gloss, a written form of ASL. Second, we pass the output to the WebSign Plug-in to play the sign. Contributions of this work are the use of a new couple of language English/ASL and an improvement of statistical machine translation based on string matching thanks to Jaro-distance.

Keywords: *Sign Language Processing, Machine Translation, Jaro Distance, Natural Language Processing.*

1. Introduction

For many centuries, Deaf have been ignored, considered mentally ill. And there wasn't effort to try to contact them. Only close deaf communicated with each other. In the 18th century, deaf people are beginning the use of a Sign Language (SL) based on gestural actions. Gestures that can express human thought as much as a spoken language. This gestural language was not a real methodical language what follows an anarchic development of sign language for a long time. Within seventies that hearing persons wishing to learn the language of the deaf and the deaf willing to teach find themselves in the school to learn. It is therefore necessary to develop teaching materials and accessible educational tools. It is very unfortunate but there is no universal sign language, each country has its own sign language. Communication between the deaf and hard of hearing from different countries or community is a problem, knowing that most deaf people do not know how to read or write. From the Eighties, researchers begin to analyze and process sign language. Next, they design and develop routines for communication intra-deaf and between hearing and deaf people. Starting from the design of automatic annotation system of the various components of sign language and coming to the 3D synthesis of signs through virtual avatars. In recent years, there was the appearance of a new line of research said automatic Sign Language Processing noted SLP. SLP is how to design, represent and process sign language incompletely described [1]. After that, there was the appearance of some works towards translate automatically written text to sign language. There are two types of machine translation. First, those which generate a 3D synthesis through a virtual character who plays the role of an interpreter in sign language. Second, those generate glosses from written text. Usually, any automatic processing of language (natural or signed) requires corpus to improve treatment outcomes. Note that sign languages are made up of manuals components and non-manual components such as gaze, facial expressions and emotions. The purpose of this paper is to focus on how to use statistics to implement a machine translation for sign language. This paper begins with an overview of various kind of machine translation for sign language. Next, an overview of our contribution and structure is introduced. Section 4 is a short description of the parallel data English/ASL. Alignment and training steps are shown in section 5. Section 6 describes our contribution in statistical machine translation. Phrase-base model and decoding are explained in section 7. Results and word alignment matrix are illustrated in section 8. Conclusions are described in section 9.

2. Machine Translators to Sign Language

Machine translators have become more reliable and effective through the development of methodology for calculating and computing power. The first translators appeared in the sixties to translate Russian into English. The first machine translator considered this task as a phase of encryption and decryption. Today and following technological developments, there was appearance of new systems. Some based on grammatical rules and other based on statistics. Not forgetting that there are translators which are based on examples. The translation stage requires preprocessing of the source language as sentence boundary detection, word tokenization, chunking… And, these treatments requires corpus. After the evolution of corpora size of and diversity for written language, there were multitudes of machine translators for the majority of languages in the world. But for sign language, we found only a few projects that translate a textual language to sign language or sign language to written text. In what follows we present various existing projects for sign language.

2.1 TEAM Project

TEAM [2] was an English-ASL translation system. It built at the University of Pennsylvania that employed synchronous tree adjoining grammar rules to construct an ASL syntactic structure. The output of the linguistic portion of the system was a written ASL gloss notation system [3] with embedded parameters. This notation system encoded limited information about morphological variations, facial expressions, and sentence mood. For synthesis, the authors took advantage of the virtual human modeling research by using an animated virtual character as signing avatar. The project had particular success at generating aspectual and adverbial information in ASL using emotive capabilities of the animated character.

2.2 English to American Sign Language: Machine Translation of Weather reports

As is common with machine translation systems, the application [4] consists of four components: a lexical analyzer, a parser, a transfer module and a generation module. In addition, there is an initial module that obtains the weather reports from the World Wide Web. Several of the components use freely available Perl modules, packages designed to assist in those particular tasks for spoken or computer languages. The ASL generation module uses the notion of “sentence stems” to generate fluent ASL. The Perl script first takes an inventory of the kinds of information present in the semantic representation, and generates a formulaic phrase for each one. These formulas all use ASL grammar, including topic-comment structure and non-manual grammatical morphemes. The content that is output by the transfer module is then plugged in to the formulas, producing fluent ASL.

2.3 The South African Sign Language Machine Translation

The aim of the South African Sign Language Machine Translation (SASL-MT) project [5] is to increase the access of the Deaf community to information by developing a machine translation system from English text to SASL for specific domains where the need is greatest, such as clinics, hospitals and police stations, providing free access to SASL linguistic data and developing tools to assist hearing students to acquire SASL. The system reuses the same concept of TEAM Project [2]. So, authors constructed SASL grammar rules, and rule-based transfer rules from the English trees to SASL trees. These trees were built manually from a set of sentences. The system transferred all pronouns detected in the sentence to objects. Then, it placed them into signing space.

This project is still under development. The authors have completed the tag parser for the English, the metadata generator for pronoun resolution and generation of emotional, stress and accent flags, and the signing avatar. Also, there aren’t experimental results.

2.4 Multipath-architecture for SL MT

Huenfaurth [6] described a new semantic representation that uses virtual reality scene. The aim of his work was to produce spatially complex American Sign Language (ASL) phenomena called “Classifier Predicates” [7]. The model acted as an Interlingua within new multi-pathway machine translation architecture. As opposed to spoken and written languages, American Sign Language relied on the multiple simultaneous channels of hand shape, hand location, hand/arm movement, facial expression and other non-manual gestures to convey the meaning. For this reason, the author used a multi-channel architecture to express additional meaning of ASL.

2.5 Czech Sign Language Machine Translation

The goal of this project was to translate spoken Czech to Signed Czech [8]. The system included a synthesis of sign by the computer animation. The synthesis employed a symbolic notation HamNoSys [9]. An automatic process of synthesis generated the articulation of hands from the notation. The translation system has built in the statistical ground. The inventory of Czech Sign Language used for the creating of complete vocabulary of signs. This dictionary had more than 3000 simple or linked signs and covers the fundamental vocabulary of Czech Deaf community.

2.6 ViSiCAST Translator

Marshall et al. at the University of East Anglia implemented a system for translating from English text into British Sign Language (BSL) [10]. Their approach used the CMU Link Parser to analyze an input English text. And they used Prolog declarative clause grammar rules to convert this linkage output into a Discourse Representation Structure. During the generation half of the translation process, Head Driven Phrase Structure rules are used to produce a symbolic SL representation script. This script is in the system’s proprietary ‘Signing Gesture Markup Language (SiGML)’, a symbolic coding scheme for the movements required to perform a natural Sign Language.

2.7 ZARDOZ System

The ZARDOZ system [11] was a proposed English-to-Sign-Languages translation system using a set of hand-coded schemata as an Interlingua for a translation component. Some of the researches focused of this system were the use of artificial intelligence knowledge representation, metaphorical reasoning, and blackboard system architecture; so, the translation design is very knowledge and reasoning heavy. During the analysis stage, English text would undergo sophisticated idiomatic concept decomposition before syntactic parsing in order to fill slots of particular concept/event/situation schemata. The advantage of the logical propositions and labeled slots provided by a schemata-architecture was that commonsense and other reasoning components in the system could later easily operate on the semantic information.

2.8 Environment for Greek Sign Language Synthesis

The authors [12] present a system that performs SL synthesis in the framework of an educational platform for young deaf children. The proposed architecture is based on standardized virtual character animation concepts for the synthesis of sign sequences and lexicon-grammatical processing of Greek sign language (GSL) sequences. A major advantage of the proposed architecture is that it goes beyond the usual single-word approach which is linguistically incorrect, to provide tools to dynamically construct new sign representations from similar ones. Words and phrases are being processed and the resulting notation subset of a lexical database, HamNoSys [9] eventually transformed into GSL and animated on the clients' side via an H|Anim compliant avatar.

2.9 Thai - Thai Sign Machine Translation

The authors [13] propose a multi-phase approach, Thai-Thai Sign Machine Translation (TTSMT), to translate Thai text into Thai Sign language. TTSMT begins the translation process by segmenting the input sentence since Thai is a non-word boundary language, converting the segmented sentence into simple sentence forms since most Thai Sign are expressed in a sequence of such form, and then generating the intermediate sign codes which link a Thai word to its corresponding Thai Sign. The most appro-priate sign codes will be selected and rearranged in the spatial grammatical order for generating the Sign language with pictures. The distinction between the Thai text and Thai Sign Language in both grammar and vocabulary are concerned in each processing step to ensure the accuracy of translation. The developed system was implemented and tested to translate Thai sentences used in everyday life.

In this section, we talked about several projects aiming to translate written text to sign language. In what follows, we introduce our contribution.

3. Contribution and structure

3.1 Problematic

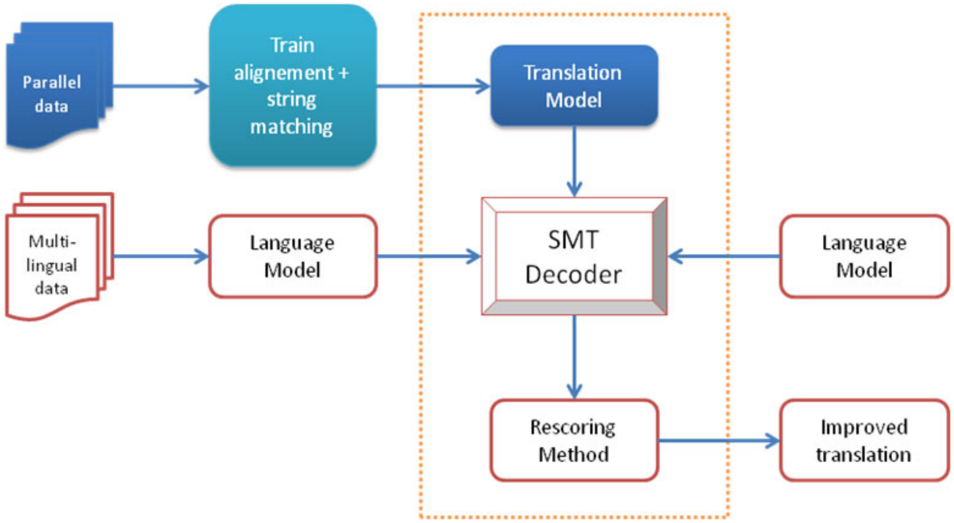
American Sign Language has emerged as the most structured Sign Language in the World. More than 20 countries, that their deaf communities sign ASL. In USA, the community of Deaf counts between one and two millions that uses ASL for communication. Deaf people can't read or write English. This is the main problem in their life. Nowadays, Internet and any tool for communication are very important in our life. So, they are not accessible for Deaf. This work aims to design a machine translation for the pair English/ASL toward helping Deaf people. It will be very helpful for interpreters, hearing people and Deaf education.

3.2 Approach

Figure 1 describes the full process of our statistical machine translation for sign language between English /ASL. In the beginning we prepare our parallel data. In fact, this data is a simple file that contains a pair of sentences, one in English and the second one in ASL that is described in the next section. This pipeline is inspired from the work of Koehn, Och and Marcu [14]. For word alignment, we used the GIZA++ statistical word alignment toolkit. This tool extracted a set of high-quality word alignment from the original unidirectional alignments sets. We include in this step a string matching algorithm. For Statistical Machine Translation (SMT) Decoder, we use MOSES [15].

4. Parallel Data: Sign Language Corpus

A corpus is a scientifically prepared collection of examples of how a language is commonly used. A corpus can contain a large number of written texts, or recorded or filmed conversations. Such data collections are used to explore the usage of a language or to find out about the vocabulary and grammar of this language. Sign Language is characterized by its interactivity and multimodality, which cause difficulties in data collection and annotation. Our corpus is composed by a pair of sentences English vs. American Sign Language (ASL). They are stored in a text file. ASL is annotated by gloss. Glosses [3] are written words, where one gloss represents one sign. Additional markings provide further information, e.g. non-manual signs. Unfortunately no gloss standard exists, which results in inconsistent annotated corpora. Figure 2 is a short dialogue between two deaf peoples. The conversation is stored into a file and ready for training.



“Figure 1**.** Statistical Sign Language Machine Translation Pipeline”

A: (get-attention) TOMORROW I GO PICK-up BOOK NEW I BUY YOU DON'T-MIND I BORROW YOUR TRUCK?
“Tomorrow I'm going to pick up some new books I just bought. Do you mind if I borrow your truck?”
B: TOMORROW TIME WHAT?
“What time tomorrow?”
A: AROUND 10 “give-or-take”
“Maybe around 10.”
B: NO NOT WORK MY TRUCK me-BRING MECHANIC FIX TOMORROW MORNING. TOMORROW AFTERNOON BETTER.
“No, that won't work; I need to take the truck to get serviced tomorrow morning. The afternoon would work better.”
A: FINE. YOU 2 TOMORROW FINE ?
“That's fine. Would 2 work for you?”
B: SURE-SURE FINE.
“Yes, that works fine.”

Figure 2. Example of conversation between 2 deaf peoples. Bold text is the ASL gloss and italic text is the English written version

5. Alignment and training

5.1. Word-based models

We present in this section a simple model for sign language machine translation that is based on lexical translation, the translation of words [15]. This method requires a dictionary that maps words from source language to target language, for example, from English to American Sign Language (ASL). If we take the word ‘*your*’, we may find multiple translations to ASL like ‘YOUR’ or ‘YOU’. Most words have multiple translations and some are more likely than others. For this reason, in some case, we cannot find the best translation if we use a dictionary to translate a sentence or a text. We refer to the use of statistics based on the count of words in a corpus or bilingual corpus. Table 1 displays the possible outcome of the word ‘your’. This word occurs 148 times in our hypothetical text collection. It is translated 119 times into ‘YOU’ and 29 times in ‘YOUR’, and so on if there are other possible translations.

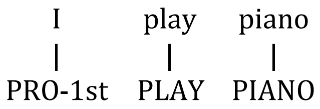
According to Koehn [15], we put formally the estimation of the lexical translation probability distribution from these counts. This function will returns a probability, for each choice of ASL translation e , that indicates how likely that translation is.

$$P_f = e \rightarrow P_f(e)$$

Table 1. Hypothetical counts for different translations of the English word ‘your’

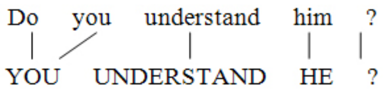
Translation of ‘your’	Count
YOU	119
YOUR	29

Thanks to probability distribution for lexical translation, we can make a leap to our first model of statistical sign language machine translation, which use only lexical translation probabilities. We denote the probability of translating an English word f into an ASL word e with the conditional probability function $t(e|f)$. The alignment between input words and output words can be illustrated by a diagram:

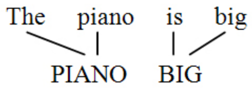


An alignment can be formalized with an alignment function a. This function maps, in our example, each ASL output word at position i to an English input word at position j: $a: j \rightarrow i$

This is a very simple alignment, since the English words and their ASL counterparts are in exactly the same order. While many languages do indeed have similar word order, a foreign language may have sentences in a different word order than is possible in ASL. This means that words have to be reordered during translation, as the following example illustrates:



$a: \{1 \rightarrow 1; 1 \rightarrow 2; 2 \rightarrow 3; 3 \rightarrow 4; 4 \rightarrow 5\}$



$a: \{1 \rightarrow 1; 1 \rightarrow 2; 2 \rightarrow 3; 2 \rightarrow 4\}$

We have just laid some examples for alignment model based on words. Note that, in our alignment model, each output can be linked with one or more than input words, as defined by the alignment function. Several works implement this model, for example, the IBM Model for word alignment that is based on lexical translation probabilities [16]. There is 5 IBM Model for mapping words from a source language and a target language. For sign language machine translation and through the experimental results, we will implement only the three first models with an improvement algorithm based on string matching. In what follows, an implementation of three IBM Model is described.

5.2 IBM Model 1, 2 and 3

IBM Model 1 defines the translation probability for an English sentence $f = (f_1, \ldots, f_{lf})$ of length l_f to an ASL sentence $e = (e_1, \ldots, e_{le})$ of length l_e with an alignment of each ASL word e_j to an English word f_i according to the alignment function $a: j \rightarrow i$ as follows:

$$p(e,a|f) = \frac{\epsilon}{(l_f + 1)^{l_e}} \prod_{j=1}^{l_e} t(e_j|f_{a(j)})$$

Let us take a look at how the algorithm words on a simple example. Table 2 presents a few iterations on a tiny three-sentence corpus with four input words (i, understand, play, piano) and four output words (I, UNDERSTAND, PLAY, PIANO). Initially, the translation probability distributions from the English words to the ASL words are ¼=0.25. Given this initial model, we collect counts in the first iteration of the EM algorithm. All alignments are equally likely.

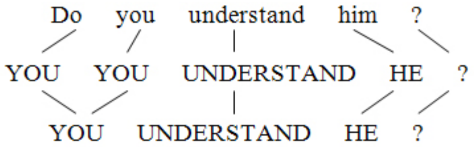
Table 2. Application of IBM Model 1 EM Training: Given the three sentence pairs, the algorithm converges to values for t(e|f)

e	f	initial	1st it.	2nd it.	3rd it.	..	Final
I	i	0.25	0.41	0.53	0.64	..	1.0
I	understand	0.25	0.50	0.45	0.38	..	0.0
I	play	0.25	0.33	0.27	0.23	..	0.0
I	piano	0.25	0.33	0.27	0.23	..	0.0
PIANO	i	0.25	0.16	0.12	0.09	..	0.0
PIANO	play	0.25	0.33	0.36	0.38	..	0.5
PIANO	piano	0.25	0.33	0.36	0.38	..	0.5
PLAY	i	0.25	0.16	0.12	0.09	..	0.0
PLAY	play	0.25	0.33	0.36	0.38	..	0.5
PLAY	piano	0.25	0.33	0.36	0.38	..	0.5
UNDERSTAND	i	0.25	0.25	0.21	0.17	..	0.0
UNDERSTAND	understand	0.25	0.50	0.55	0.61	..	1.0

In IBM Model 2, we add an explicit model for alignment. In IBM Model 1, we do not have a probabilistic model for this aspect of translation. As consequence, according to IBM Model 1 the translation probabilities for the two examples cited previously are the same. IBM Model 2 addresses the issue of alignment with an explicit model for alignment based on the positions of the input and output words. The translation of an English input word in position i to an ASL word in position j is modeled by an alignment probability distribution:

$$a(i|j, l_e, l_f)$$

Recall that the length of the input sentence f is denoted as l_f, and the length of the output sentence e is l_e. We can view translation under IBM Model 2 as a two-step process with a lexical translation step and an alignment step:



The first step is lexical translation as in IBM Model 1, again modeled by the translation probability t(e|f). The second step is the alignment step. For instance, translating 'understand' into 'UNDERSTAND' has a lexical translation probability of :

$$\mathrm{t(UNDERSTAND\,|\,understand)}$$

and an alignment probability of a(2|3,4,5) - the 2th ASL word is aligned to the 3rd English word.

Note that the alignment function a maps each ASL output word j to an English input position a(j) and the alignment probability distribution is also set up in this reverse direction. The two steps are combined mathematically to form IBM Model 2:

$$p(e,a|f) = \epsilon \prod_{j=1}^{l_e} t\left(e_j|f_{a(j)}\right) a\left(a(j)|j,l_e,l_f\right)$$

Table 3 shows that in only three iterations we achieve the same results of IBM Model 1.

In IBM Model 3, we account the NULL token. In other words, you can get a word in English that is not translated into ASL. The probability of generating a NULL token is:

$$p(\emptyset_0) = \binom{l_e - \emptyset_0}{\emptyset_0} p_1^{\emptyset_0} p_0^{l_e - 2\emptyset_0}$$

Table 3. Application of IBM Model 2 EM Training

e	f	initial	1st it.	2nd it.	3rd it.
I	i	0.64	0.73	0.96	1.00
I	understand	0.38	0.32	0.09	0.00
I	play	0.23	0.18	0.03	0.00
I	piano	0.23	0.18	0.03	0.00
PIANO	i	0.09	0.06	0.00	0.00
PIANO	play	0.38	0.40	0.48	0.50
PIANO	piano	0.38	0.40	0.48	0.50
PLAY	i	0.09	0.06	0.00	0.00
PLAY	play	0.38	0.40	0.48	0.50
PLAY	piano	0.38	0.40	0.48	0.50
UNDERSTAND	i	0.17	0.13	0.01	0.00
UNDERSTAND	understand	0.61	0.67	0.90	1.00

Due to the problem of incomplete and according to Koehn, we are facing a typical problem for machine learning. We want to estimate our model from incomplete data. So, we will use the expectation Maximization algorithm, or EM Algorithm that addresses the situation of incomplete data. It is an iterative learning method that fills in the gaps in the data and trains a model in alternating steps. We apply EM for IBM Model 1, 2 and 3.

6. String Matching

Words in American Sign Language are very similar to English written text. So, we think to use others techniques to learn data quickly and efficiency, for example, string-matching. String-matching is a very important subject in the wider domain of text processing. String-matching algorithms are basic components used in implementations of practical software existing under most operating systems. Moreover, they emphasize programming methods that serve as paradigms in other fields of computer science. They also play an important role in theoretical computer science by providing challenging problems. String-matching consists in finding one, or more generally, all the occurrences of a string in a text or with another string. The pattern is denoted by $x = x[0\,..\,m-1]$; its length is equal to m. The text is denoted by $y = y[0\,..\,n-1]$; its length is equal to n. Both strings are build over a finite set of character called an alphabet denoted by with size is equal to. Several algorithms and methods exist like Jaro-Winkler distance that does will be used in word alignment process from statistical sign language machine translation.

6.1 Jaro-Winkler distance

The Jaro–Winkler distance [17] is a measure of similarity between two strings. It is a variant of the Jaro distance metric and mainly used in the area of record linkage. The higher the Jaro–Winkler distance for two strings is, the more similar the strings are. The Jaro–Winkler distance metric is designed and best suited for short strings such as person names. The score is normalized such that 0 equates to no similarity and 1 is an exact match. The Jaro distance d_j of two given strings S_1 and S_2 is:

$$d_j = \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right)$$

Where:

- m is the number of matching characters
- t is the number of transpositions

Jaro–Winkler distance uses a prefix scale p which gives more favorable ratings to strings that match from the beginning for a set prefix length l. Given two strings S_1 and S_2, their Jaro–Winkler distance d_w is: $d_w = d_j + \left(l.p.(1-d_j)\right)$

Where:

- d_j is the Jaro distance between s_1 and s_2.
- l is the length of common prefix at the start of the string up to maximum of 4 characters.
- p is constant scaling factor for how much the score is adjusted upwards for having common prefixes. p should not exceed 0.25, otherwise the distance can become larger than 1. The standard value for this constant in Winkler's work is p = 0.1.

The next Table presents some examples:

Table 4. Jaro-Winkler distance applied to 5 pairs word

S1	S2	Jaro distance	Jaro-Winkeler distance
I	i	1.00	1.0000
I	understand	0.00	0.0000
PIANO	play	0.38	0.4550
PIANO	piano	1.00	1.0000
UNDERSTAND	understand	1.00	1.0000

6.2 String Matching for EM for IBM Model 1

Starting from the formula: $p(a|e,f) = \frac{p(e,a|f)}{p(e|f)}$

We improve the result by adding the d_w between e and f, we have:

$$p(a|e,f) = \frac{\alpha.\,p(e,a|f) + (1-\alpha).\,d_w(e,f)}{p(e|f)}$$

Where α is the coefficient of similiraty between the two words e and f. The standard value of αused for experiments is 0.5. Table 2 presents comparative results applied to a small corpus composed by two pair-sentences.

Table 5. Application of IBM Model 1 EM Training with string matching

e	**f**	**initial**	**3 iterations**	**3 iterations + String Matching**
I	i	0.2500	0.6412	0.9684
I	understand	0.2500	0.3879	0.0532
I	play	0.2500	0.2307	0.0316
I	piano	0.2500	0.2307	0.0839
PIANO	i	0.2500	0.0929	0.0475
PIANO	play	0.2500	0.3846	0.3738
PIANO	piano	0.2500	0.3846	0.7977
PLAY	i	0.2500	0.0929	0.0049
PLAY	play	0.2500	0.3846	0.8170
PLAY	piano	0.2500	0.3846	0.3741
UNDERSTAND	i	0.2500	0.1727	0.0123
UNDERSTAND	understand	0.2500	0.6120	0.9467

6.3 String Matching for EM for IBM Model 2

Table 6. Application of IBM Model 2 EM Training with string matching

e	**f**	**initial**	**2 iterations**	**2 iterations + string matching**
I	i	0.6412	0.7343	0.9986
I	understand	0.3879	0.3258	0.0030
I	play	0.2307	0.1899	0.0010
I	piano	0.2307	0.1899	0.0380
PIANO	i	0.0929	0.0657	0.0345
PIANO	play	0.3846	0.4050	0.3169
PIANO	piano	0.3846	0.4050	0.9047
PLAY	i	0.0929	0.0657	0.0000
PLAY	play	0.3846	0.4050	0.9044
PLAY	piano	0.3846	0.4050	0.3129
UNDERSTAND	i	0.1727	0.1341	0.0001
UNDERSTAND	understand	0.6120	0.6741	0.9969

Like to IBM Model 1, we add an α coefficient for string matching to alignment process. Results show that we converge to 1 after 2 iterations only in a small corpus. We note that the corpus contains two similar words but have not the same semantic and role ‘piano’ and ‘play’. The next table presents the experimental results.

7. Phrase-based model and Decoding

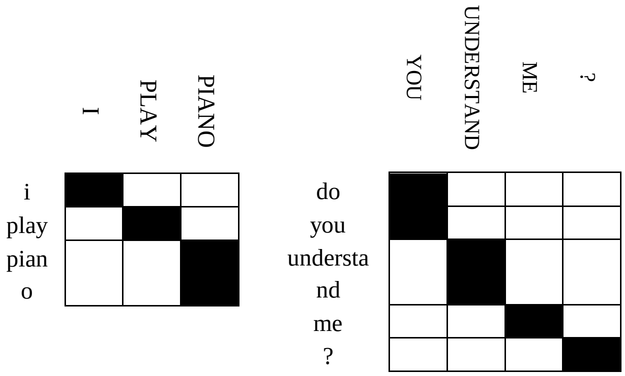
In phrase-based translation, the aim is to reduce the restrictions of word-based translation by translating whole sequences of words, where the lengths may differ. We use MOSES tool to learn phrase alignment. After that, we exploit the decoding tool. This step is the main function in the system. The input is an English sentence. The role of the decoder is to find the best translation. The probabilistic model for phrase-based translation is:

$$e_{max} = argmax_e \prod_{i=1}^{l} \emptyset(\overline{f_i}|\overline{e_i}) d(start_i - end_{i-1} - 1) P_{LM}(e)$$

Where:

- Phrase translation: picking phrase $\overline{f_i}$ to be translated as a phrase $\overline{e_i}$. We look up score $\emptyset(\overline{f_i}|\overline{e_i})$ from phrase translation table.
- Reordering: Previous phrase ended in end_{i-1}, current phrase starts at $start_i$. We compute $d(start_i - end_{i-1} - 1)$.
- Language Model : For n-gram model, we need to keep track of last $(n-1)$ words. We compute score $P_{LM}(W_i|W_{i-(n-1)}, \dots, W_{i-1})$ for added words W_i.

8. Results and Word alignment matrix



	I	PLAY	PIANO
i	■		
play		■	
pian o			■

	YOU	UNDERSTAND	ME	?
do	■			
you	■			
understa nd		■		
me			■	
?				■

Figure 3. Word alignment matrix: Words in the ASL (columns) are aligned to words in the English sentence (rows) as indicated by the filled points in the matrix

One way to visualize the task of word alignment is by a matrix as in Figure 3. Here, alignments between words (for instance between the English 'play' and the ASL 'PIANO') are represented by points in the alignment matrix. Word alignments do not have to be one-to-one. Words may have multiple or no alignment points. For instance, the ASL word assumes is aligned to the two English words 'do you'. However, it is not always easy to establish what the correct word alignment should be. Experimentation and results

In this evaluation, we trained a small 3-gram language model using data in Table 7.

Results are very encouraged. Table 8 shows some alignment sentences with scores.

For interpretation, we use WebSign tool [18]. WebSign is a project that carries on developing tools able to make information over the web accessible for deaf.

Table 7. Statistics of Parallel data

Language	Sentences	Tokens
English	431	632
ASL	431	608
n-gram 1 = 609 - n-gram 2 = 1539 - n-gram 3 = 257		

9. Conclusions

We describe several experiments with English-to-American Sign Language statistical sign language machine translation. Employing a technique of string matching is crucial. In conclusion, phrase-based statistical MT for sign language from English to American Sign Language performs well, despite the expectations arising from linguistic knowledge about the properties of ASL. This work we experimented with is currently the best performing machine translation evaluated on this pair of languages.

References

[1] P. Dreuw, D. Stein, T. Deselaers, D. Rybach, M. Zahedi, J. Bungeroth, H. Ney, "Spoken Language Processing Techniques for Sign Language Recognition and Translation", Technology and Disability, Vol. 20(2), 2008, pp. 121-133.

[2] L. Zhao, K. Kipper, W. Schuler, C. Vogler, N. Badler, M. Palmer, "A Machine Translation System from English to American Sign Language", Envisioning Machine Translation in the Information Future, Vol. 1934, 2000, pp. 191-193.

[3] K. Liddell, Grammar, gesture, and meaning in American Sign Language, Cambridge University Press, 2003.

[4] A. Grieve-Smith, "English to American Sign Language Machine Translation of Weather Reports", Proceedings of the Second High Desert Student Conference in Linguistics, 1999, pp. 23–30.

[5] L. Van Zijl, D. Barker, "South African Sign Language Machine Translation System", Proceedings of the Second International Conference on Computer Graphics, Virtual Reality, Visualization and Interaction in Africa, 2003, pp. 49-52.

[6] M. Huenerfauth, "American Sign Language Generation: Multimodal NLG with Multiple Linguistic Channels", Proceedings of the ACL Student Research Workshop, 2005, pp. 37–42.

[7] M. Huenerfauth, "Spatial and Planning Models of ASL Classifier Predicates for Machine Translation", The 10th International Conference on Theoretical and Methodological Issues in Machine Translation, 2004, pp. 65–74.

[8] J. Kanis, J. Zahradil, F. Jurčíček, L. Müller, "Czech-Sign Speech Corpus for Semantic based Machine Translation", Text, Speech and Dialogue, Vol. 4188, 2006, pp. 613-620.

[9] T. Hanke, "HamNoSys – Representing Sign Language Data in Language Resources and Language Processing Contexts", LREC Representation and processing of sign languages, 2004, pp. 1-6.

[10] J. A. Bangham, S. J. Cox, R. Elliott, J. R. W. Glauert, I. Marshall, S. Rankov, M. Wells, "Virtual Signing: Capture, Animation, Storage and Transmission – an Overview of the

ViSiCAST Project", Speech and Language Processing for Disabled and Elderly People, 2000.

[11] T. Veale, A. Conway, "Cross modal comprehension in ZARDOZ an English to sign-language translation system", INLG '94 Proceedings of the Seventh International Workshop on Natural Language Generation, 1994, pp. 249-252.

[12] K. Karpouzis, G. Caridakis, S. Fotinea, E. Efthimiou, "Educational resources and implementation of a Greek sign language synthesis architecture", Web3D Technologies in Learning, Education and Training 49, 2007, pp. 54-74.

[13] S. Dangsaart, K. Naruedomkul, N. Cercone, B. Sirinaovakul, "Intelligent Thai text – Thai sign translation for language learning", Computers & Education, 2008, pp. 1125-1141.

[14] P. Koehn, F. Och, D. Marcu, "Statistical phrase-based translation", NAACL'03 Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology 1, 2003, pp. 48-54.

[15] P. Koehn, Statistical Machine Translation, Cambridge University Press, 2009.

[16] P. Brown, V. Pietra, S. Pietra, R. Mercer, "The mathematics of statistical machine translation: parameter estimation", Computational Linguistics - Special issue on using large corpora: II, Vol. 19, 1993, pp. 263-311.

[17] M. Jaro, "Advances in record-linkage methodology as applied to matching the 1985 census of Tampa", Journal of the American Statistical Association, 1989, pp. 414–420.

[18] M. Jemni, O. El Ghoul, "An avatar based approach for automatic interpretation of text to Sign language", 9th European Conference for the Advancement of the Assistive Technologies in Europe, 2007.

Achraf Othman is currently a PhD student under the supervision of Prof. Mohamed Jemni. He received in August 2010 the Master degree on Computer Science from Tunis College of Sciences and Techniques (ESSTT), University of Tunis in Tunisia. His research interests are in the areas of Sign Language Processing. His current topics of interests include Grid Computing, Computer graphics and Accessibility of ICT to Persons with Disabilities.

Mohamed Jemni is a Professor of ICT and Educational Technologies at the University of Tunis, Tunisia. He is the Head of the Laboratory Research of Technologies of Information and Communication (UTIC). Since August 2008, he is the General chair of the Computing Center El Khawarizmi, the internet services provider for the sector of the higher education and scientific research. His Research Projects Involvement are tools and environments of e-learning, Accessibility of ICT to Persons with Disabilities and Parallel & Grid Computing.

Table 8. Some alignments sentences (English / American Sign Language) with scores

ASL Sentence pair **English sentence + alignment**	**Length :** **Source / Target**	**Score**
DEAF YOU ? NULL ({ }) are ({ }) you ({ 2 }) deaf ({ 1 }) ? ({ 3 })	Source : 4 Target : 3	0.0016781
YOU UNDERSTAND SHE , YOU ? NULL ({ 4 }) do ({ 5 }) you ({ 1 }) understand ({ 2 }) her ({ 3 }) ? ({ 6 })	Source : 5 Target : 6	5.387e-07
YOU FAVORITE- [prefer] , HAMBURGER [body-shift-or] HOTDOG ? NULL ({ 6 }) do ({ }) you ({ 1 }) prefer ({ 2 5 10 }) hamburgers ({ 7 9 }) or ({ 3 8 }) hotdogs ({ 4 11 }) ? ({ 12 })	Source : 7 Target : 12	2.195e-16
last-YEAR TICKET HOW-MANY YOU ? NULL ({ }) how ({ }) many ({ 3 }) tickets ({ 1 2 }) did ({ }) you ({ 4 }) get ({ }) last ({ }) year ({ }) ? ({ 5 })	Source : 9 Target : 5	3.661e-06
TOPIC YOU DON 'T-CARE WHAT ? NULL ({ }) what ({ }) do ({ }) you ({ 2 }) not ({ 3 }) care ({ 1 4 }) about ({ 5 }) ? ({ 6 })	Source : 7 Target : 6	5.444e-06
DRESS YOU LIKE USE- [wear] YOU ? NULL ({ }) do ({ 8 }) you ({ 2 }) like ({ 3 }) to ({ }) wear ({ 1 4 6 }) dresses ({ 5 7 }) ? ({ 9 })	Source : 8 Target : 7	2.799e-11
WET-WIPES YOU KEEP CAR ? NULL ({ }) do ({ }) you ({ 2 }) keep ({ }) wet ({ 1 }) wipes ({ 3 }) in ({ }) your ({ }) car ({ 4 }) ? ({ 5 })	Source : 9 Target : 5	4.286e-05
STRIPES- [vertical] , YOU FACE- [look] GOOD YOU ? NULL ({ }) do ({ 6 }) you ({ 12 }) look ({ 4 7 10 }) good ({ 2 8 11 }) in ({ 5 }) stripes ({ 1 3 9 }) ? ({ 13 })	Source : 7 Target : 13	6.651e-19
# Sentence pair (430) source length 6 target length 4 alignment score : WHAT YOU ENTHUSIASTIC ? NULL ({ }) what ({ }) are ({ }) you ({ 2 }) enthusiastic ({ 3 }) about ({ 1 }) ? ({ 4 })	Source : 6 Target : 4	0.0006027

Minutiae Extraction from Fingerprint Images - a Review

Roli Bansal[1], Priti Sehgal[2] and Punam Bedi[3]

**[1] Department of Computer Science,
University of Delhi, New Delhi - 110001, India.**

**[2] Reader, Department of Computer Science, Keshav Mahavidyalaya,
University of delhi, Pitampura , New Delhi - 110034, India.**

**[3] Associate Professor, Department of Computer Science,
University of Delhi, New Delhi - 110001, India.**

Abstract

Fingerprints are the oldest and most widely used form of biometric identification. Everyone is known to have unique, immutable fingerprints. As most Automatic Fingerprint Recognition Systems are based on local ridge features known as minutiae, marking minutiae accurately and rejecting false ones is very important. However, fingerprint images get degraded and corrupted due to variations in skin and impression conditions. Thus, image enhancement techniques are employed prior to minutiae extraction. A critical step in automatic fingerprint matching is to reliably extract minutiae from the input fingerprint images. This paper presents a review of a large number of techniques present in the literature for extracting fingerprint minutiae. The techniques are broadly classified as those working on binarized images and those that work on gray scale images directly.

Keywords: *fingerprint images, minutiae extraction, ridge endings, ridge bifurcation, fingerprint recognition.*

1. Introduction

Biometrics is the science of uniquely recognizing humans based upon one or more intrinsic physical or behavioral traits. Fingerprints are the most widely used parameter for personal identification amongst all biometrics. Fingerprint identification is commonly employed in forensic science to aid criminal investigations etc. A fingerprint is a unique pattern of ridges and valleys on the surface of a finger of an individual. A ridge is defined as a single curved segment, and a valley is the region between two adjacent ridges. Minutiae points (fig. 1) are the local ridge discontinuities, which are of two types: ridge endings and bifurcations. A good quality image has around 40 to 100 minutiae [1]. It is these minutiae points which are used for determining uniqueness of a fingerprint.

Automated fingerprint recognition and self authentication systems [2] can be categorized as verification or identification systems.

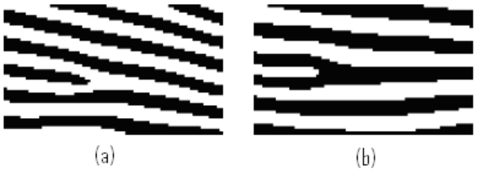


Fig. 1 Minutiae Points. (a) ridge ending (b) bifurcation

The verification process either accepts or rejects the user’s identity by matching against an existing fingerprint database. In identification, the identity of the user is established using fingerprints. Since accurate matching of fingerprints depends largely on ridge structures, the quality of the fingerprint image is of critical importance. However, in practice, a fingerprint image may not always be well defined due to elements of noise that corrupt the clarity of the ridge structures. This corruption may occur due to variations in skin and impression conditions such as scars, humidity, dirt, and non-uniform contact with the fingerprint capture device. Many algorithms [3-14] have been proposed in the literature for minutia analysis and fingerprint matching and classification for better fingerprint verification and identification. Recently, techniques [15, 16, 17, 18] have been proposed that use other features apart from minutiae for fingerprint recognition. Chen et al [15] propose to reconstruct the fingerprint’s orientation field from minutiae and utilize it in the matching stage to improve the system’s performance. Cao et al [16] have introduced two novel features to deal with non linear distortion in fingerprints. These features are the finger placement direction and the ridge compatibility. Choi et al [17] proposed to incorporate ridge features like ridge count, ridge length, ridge curvature direction and ridge type together with minutiae to increase the matching performance. Current scientific studies show that application of evolutionary algorithms may improve the performance of biometric systems significantly [19]. There are a number of instances in the literature [20, 21] where evolutionary algorithms are used for matching minutiae of a fingerprint with that of a database of fingerprint images. The results of all such techniques depend on the quality of the input image. Thus, image enhancement techniques are often employed to reduce the noise and to enhance the definition of ridges against valleys so that no spurious minutiae are identified. In fact, matching latent fingerprints from crime scenes is difficult because of their poor quality and the fingerprint matching accuracy is improved by combining manually marked minutiae with automatically extracted ones [22]. Several methods have been proposed for enhancement of fingerprint images which are based on image normalization and Gabor filtering (Hong's algorithm) [1], Directional Fourier filtering [23], Binarization Method [24], enhancement using directional median filter[25], fingerprint image enhancement using filtering techniques[26], image retrieval based on color histogram and textual features[27] and many others[28-32]. The Hong's algorithm inputs a fingerprint image and applies various steps for enhancement. Several other enhancement techniques present in literature are based on fuzzy logic and neural networks [33-40]. Choonwoo et al [41] presented a novel approach to enhance feature extraction for low quality fingerprint images using stochastic resonance (SR). SR refers to a phenomenon where an appropriate amount of noise added to the original signal can increase the signal-to-noise-ratio. Experimental results show that Gaussian noise added to low quality fingerprint images enables the extraction of useful features for biometric identification. The rest of the paper is organized as follows: Section 2 discusses fingerprint features and section 3 explains fingerprint recognition. Section 4 lists the techniques available for minutiae extraction in the literature and finally, section 5 concludes the paper.

2. Fingerprint Features

Fingerprint features can be classified into three classes [1]. Level 1 features show macro details of the ridge flow shape, Level 2 features (minutiae point) are discriminative enough for recognition, and Level 3 features (pores) complement the uniqueness of Level 2 features.

2.1 Global Ridge Pattern

A fingerprint is a pattern of alternating convex skin called ridges and concave skin called valleys with a spiral-curve-like line shape (fig. 2). There are two types of ridge flows: the pseudo-parallel ridge flows and high-curvature ridge flows which are located around the core point and/or delta point(s). This representation relies on the ridge structure, global landmarks and ridge pattern characteristics. The commonly used global fingerprint features are:

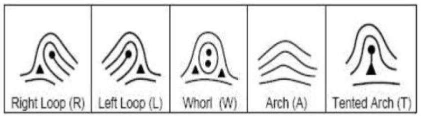


Fig. 2 Global fingerprint ridge patterns

- Singular points – They represent discontinuities in the orientation field. There are two types of singular points as shown in fig. 3. A core is the uppermost of the innermost curving ridge [1], and a delta point is the junction point where three ridge flows meet. They are usually used for fingerprint registration and classification.

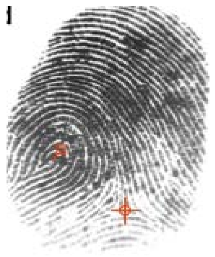


Fig. 3 Singular points (SPs), where "**O**" and "⊕" denote core and delta, respectively.

- Ridge orientation map – This represents the local direction of the ridge-valley structure. It is commonly utilized for classification, image enhancement and minutia feature verification and filtering.
- Ridge frequency map – It is the reciprocal of the ridge distance in the direction perpendicular to local ridge orientation. It is extensively utilized for contextual filtering of fingerprint images.

2.2 Local Ridge Pattern

This is the most widely used and studied fingerprint representation. Local ridge details are the discontinuities of local ridge structure referred to as minutiae. Sir Francis Galton (1822-1922) was the first person who observed the structures and permanence of minutiae. Therefore, minutiae are also called "Galton details". They are used by forensic experts to match two fingerprints. There are about 150 different types of minutiae [3] categorized based on their configuration. Among these minutia types, "ridge ending" and "ridge bifurcation" are the most commonly used, since all the other types of minutiae can be seen as

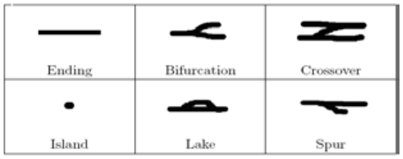


Fig. 4 Some of the common minutiae types

combinations of “ridge endings” and “ridge bifurcations”. Some minutiae are shown in fig. 4. The American National Standards Institute-National Institute of Standard and Technology (ANSI-NIST) proposed a minutiae-based fingerprint representation. It includes minutiae location and orientation [42]. Minutia orientation is defined as the direction of the underlying ridge at the minutia location (fig. 5). Minutiae-based fingerprint representation can also assist privacy issues since one cannot reconstruct the original image from using only minutiae information. Actually, minutiae are sufficient to establish fingerprint individuality.

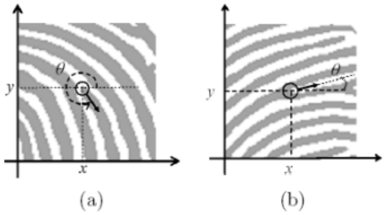


Fig. 5 (a) A ridge ending minutia: (x, y) are the minutia coordinates; θ is the minutia’s orientation; (b) A ridge bifurcation minutia: (x, y) are the minutia coordinates; θ is the minutia’s orientation.

The minutiae are relatively stable and robust to contrast, image resolutions, and global distortion as compared to other fingerprint representations. However, to extract the minutiae from a poor quality image is not an easy task, although most of the automatic fingerprint recognition systems are designed to use minutiae as the main fingerprint feature for recognition.

3 What is Fingerprint Recognition?

The fingerprint recognition [43, 44, 45] problem can be grouped into three sub-domains: fingerprint enrollment, verification and fingerprint identification. In addition, as different from the manual approach for fingerprint recognition by experts, the fingerprint recognition here is referred as AFRS (Automatic Fingerprint Recognition System), which is program-based. Verification is typically used for positive recognition, where the aim is to prevent multiple people from using the same identity. Fingerprint verification is to verify the authenticity of one person by his fingerprint. There is one-to-one comparison in this case. In the identification mode, the system recognizes an individual by searching the templates of all the users in the database for a match. Therefore, the system conducts a one to many comparisons to establish an individual’s identity. Both verification and identification use certain techniques for fingerprint matching as indicated in the following subsection.

3.1 Techniques for Fingerprint Matching

Various fingerprint matching techniques discussed in literature are as follows:

- Minutiae based technique: Most of the finger-scan technologies are based on Minutiae. Minutia based techniques represent the fingerprint by its local features, like terminations and bifurcations [46-80]. Two fingerprints match if their minutiae points match. This approach has been intensively studied, also is the backbone of the current available fingerprint recognition products.
- Pattern Matching or Ridge Feature Based Techniques: Feature extraction and template generation are based on series of ridges as opposed to discrete points which forms the basis of Pattern Matching Techniques. This includes context aware similarity search techniques applicable to all types of content based image retrieval (CBIR) [81]. The advantage of Pattern Matching techniques [82, 83] over Minutiae based techniques is that minutiae points may be affected by wear and tear and the disadvantages are that these are sensitive to proper placement of finger and need large storage for templates.
- Correlation Based Technique [84, 85, 86] : Let I(Δx, Δy, θ) represent a rotation of the input image I by an angle θ around the origin (usually the image center) and shifted by Δx and Δy pixels in directions x and y, respectively. Then the similarity between the two fingerprint images T and I can be measured as :

$$S(T,I) = \max_{\Delta x,\Delta y,\theta} CC(T, I^{(\Delta x,\Delta y,\theta)}) \qquad (1)$$

where CC(T, I) = $T^{T}\bar{I}$ is the cross-correlation between T and I. The cross-correlation is a well known measure of image similarity. It allows us to find the optimal registration. The direct application of eq. (1) rarely leads to acceptable results, mainly due to the following problems:

a) Non-linear distortion makes impressions of the same finger significantly different in terms of global structure; the use of local or block-wise correlation techniques can help to deal with this problem.

b) Skin condition and finger pressure cause image brightness, contrast, and ridge thickness to vary significantly across different impressions. The use of more sophisticated correlation measures may compensate for these problems.

c) The technique is computationally very expensive. Local correlation and correlation in the Fourier domain can improve efficiency.
- Image Based Techniques: Image based techniques try to do matching based on the global features of a whole fingerprint image. It is an advanced and newly emerging method for fingerprint recognition. It is useful to solve some intractable problems of the first approach.

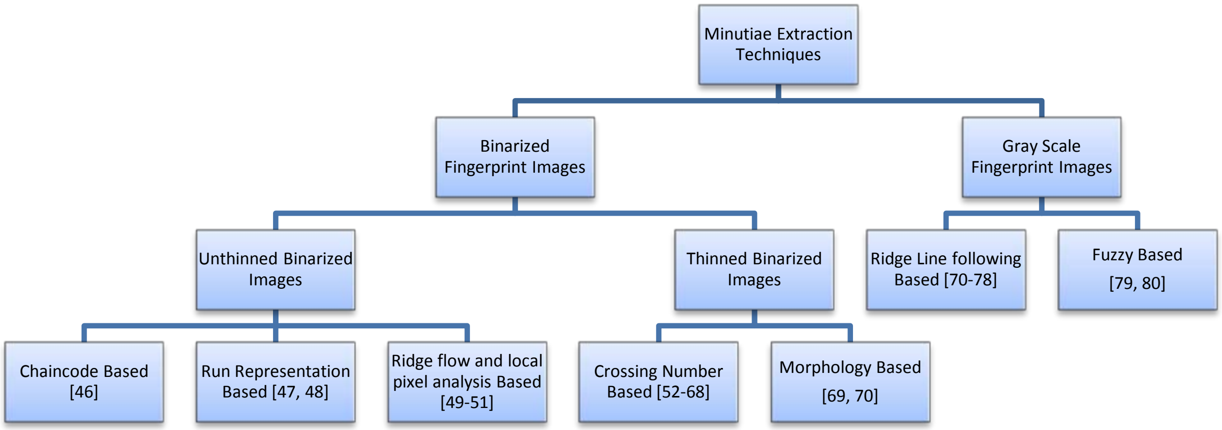

Minutiae Extraction Techniques
Binarized Fingerprint Images
Gray Scale Fingerprint Images
Unthinned Binarized Images
Thinned Binarized Images
Ridge Line following Based [70-78]
Fuzzy Based [79, 80]
Chaincode Based [46]
Run Representation Based [47, 48]
Ridge flow and local pixel analysis Based [49-51]
Crossing Number Based [52-68]
Morphology Based [69, 70]

Fig. 6 Classification of Minutiae Extraction Techniques

4. Minutiae Extraction

An accurate representation of the fingerprint image is critical to automatic fingerprint identification systems, because most deployed commercial large-scale systems are dependent on feature-based matching (correlation based techniques have problems as discussed in the previous section). Among all the fingerprint features, minutia point features with corresponding orientation maps are unique enough to discriminate amongst fingerprints robustly; the minutiae feature representation reduces the complex fingerprint recognition problem to a point pattern matching problem. In order to achieve high-accuracy minutiae with varied quality fingerprint images, segmentation algorithm needs to separate foreground from noisy background which includes all ridge-valley regions and not the background. Image enhancement algorithm needs to keep the original ridge flow pattern without altering the singularity, join broken ridges, clean artifacts between pseudo-parallel ridges, and not introduce false information. Finally minutiae detection algorithm needs to locate efficiently and accurately the minutiae points.

There are a lot of minutiae extraction methods available in the literature. We can classify these methods broadly into two categories (fig. 6):

- Those that work on binarized fingerprint images
- Those that work directly on gray-scale fingerprint images.

The following subsections elaborate on the above mentioned categories.

4.1 Minutiae detection from binarized fingerprints

A number of binary image based methods are available which detect minutiae by inspecting the localized pixel patterns. They can be further classified into two classes, those that work on unthinned binarized images and those that work on thinned binarized images.

4.1.1 Unthinned Binarized images

Most fingerprint minutia extraction methods are thinning-based where the skeletonization process converts each ridge to one pixel wide. Minutia points are detected by locating the end points and bifurcation points on the thinned ridge skeleton based on the number of neighboring pixels. The end points are selected if they have a single neighbor and the bifurcation points are selected if they have more than two neighbors. However, methods based on thinning are sensitive to noise and the skeleton structure does not conform to intuitive expectation. This category focuses on a binary image based technique of minutiae extraction without a thinning process. The main problem in the minutiae extraction method using thinning processes comes from the fact that minutiae in the skeleton image do not always correspond to true minutiae in the fingerprint image. In fact, a lot of spurious minutiae are observed because of undesired spikes, breaks, and holes. Therefore, post processing is usually adopted to avoid spurious minutiae, which is based on both statistical and structural information after feature detection. This category discusses three major techniques of minutiae extraction from unthinned binarized images based on chaincode processing [46], run based methods[47,48] and ridge flow and local pixel analysis based methods [49-51].

4.1.1.1 Chaincode processing

Chaincode representation of object contours is extensively used in document analysis. Unlike thinned skeletons, the pixel image can be fully recovered from the chaincode of its contour. In this method, the image is scanned from top to bottom and right to left. The transitions from white (background) to black (foreground) are detected. The contour is then traced counterclockwise and expressed as an array of contour elements [46]. Each contour element represents a pixel on the contour. It contains fields for the x, y coordinates of the pixel, the slope or direction of the contour into the pixel, and auxiliary information such as curvature.

In a binary fingerprint image, ridge lines are more than one pixel wide. Tracing a ridge line along its boundary in counterclockwise direction, a termination minutia (ridge ending) is detected when the trace makes a significant left turn. Similarly, a bifurcation minutia (a fork) is detected when the trace makes a significant right turn (fig. 7(a)). Let a vector P_{in} go in to a contour point P and a vector P_{out} go out of P. The computations of P_{in} and P_{out} use several neighboring contour points. This is to avoid local noise and at the same time obtain a better estimation of the vectors using the average of more than one point. The significance of the direction change at P is determined by the angle made between P_{in} and P_{out}:

$$\theta = \arccos \frac{P_{\text{in}} \cdot P_{\text{out}}}{|P_{\text{in}}| P_{\text{out}}|} \quad (2)$$

After size normalizations, let the two vectors be P_{in} = (x1,y1) and P_{out} = (x2,y2). Then,

$$\theta = \arccos(x_1y_1 + x_2y_2) \quad (3)$$

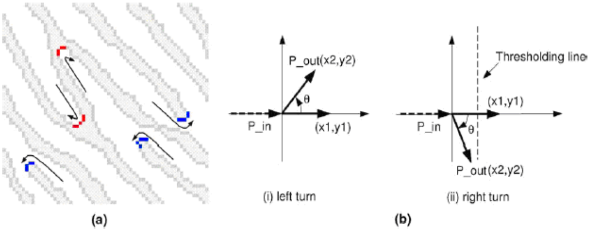


Fig. 7 (a) Minutia location in chaincode contours, the counter wise tracing along the boundary of a ridge line turns left at a termination minutia and turns left at a bifurcation minutia. (b) To calculate the significant turns, the distance between the thresholding line and the y-axis gives a threshold for determining a significant turn [46].

A threshold T is selected so that any significant turn satisfies the condition:

$$x_1y_1 + x_2y_2 < T$$

If the vectors are placed in a Cartesian coordinate system with P_{in} along the x-axis (fig. 7(b)), then the threshold T is the x-coordinate of the thresholding line. The turning direction is determined by the sign of sin h since the angle h is always in the range -90 to +90. Therefore,

$x_1y_2 - x_2y_1 > 0$ indicates a left turn and

$x_1y_2 - x_2y_1 < 0$ indicates a right turn.

This method of direction field generation using chaincode for image enhancement is more efficient and robust for the following reasons: (i) chaincode generation depends on a pre-binarization algorithm, (ii) the adaptive binarization algorithm and the chaincode generation algorithm are both efficient, and (iii) the orientation field is directly computed by tracing the chaincode over a discrete grid. The objective is to attain the ridge orientation for the entire window rather than at every pixel.

4.1.1.2 Run Representation

This method results in fast extraction of fingerprint minutiae that are based on the horizontal and vertical run-length encoding from binary images without a computationally expensive thinning process [47, 48]. Fingerprint images are represented by a cascade of runs after run-length encoding. Then runs' adjacency is checked and characteristic runs are detected. But all characteristic runs cannot be true minutiae. So, some geometric constraints are introduced for checking validity of characteristic runs. As shown in fig. 8, the image is preprocessed for enhancement, which is based on the convolution of the image with Gabor filters tuned to the local ridge orientation and frequency. Firstly, the image is segmented [87-90] to extract it from the background. Next, it is normalized so that it has a prespecified mean and variance. After calculating the local orientation and ridge frequency around each pixel, the Gabor filter is applied to each pixel location in the image. As a result the filter enhances the ridges oriented in the direction of local orientation. Hence the filter increases the contrast between the foreground ridges and the background, while effectively reducing noise to set the parameters with respect to the orientation and the frequency, respectively.

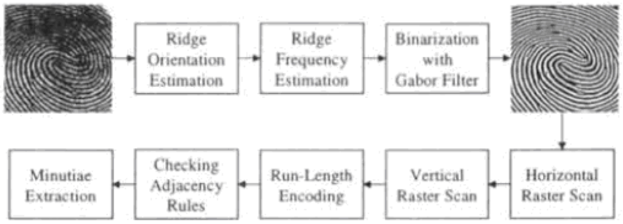


Fig. 8 Block diagram of proposed minutiae extraction algorithm using run-length encoding.

Next, the image is binarized. The simplest way to use image binarization is to choose a threshold value, and classify all pixels with values above this threshold as white, and all other pixels as black. The problem is how to select the correct threshold. In many cases, finding one threshold compatible to the entire image is very difficult, and in many cases even impossible. Therefore, adaptive image binarization is needed where an optimal threshold is chosen for each image area [91, 92].

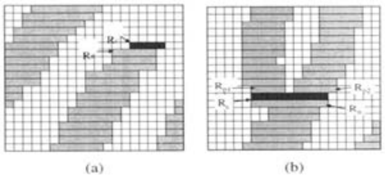


Fig. 9 The minutiae in run representation. (a) Termination in horizontal runs (b) Bifurcation in horizontal runs.

A run-length encoding is an efficient coding scheme for binary or labeled images because it can not only reduce memory space but also speed up image processing time. In the binary image, successive black pixels along the scan line are defined as a run. Generally, a run-length encoding of a binary image is a list of contiguous horizontal runs of black-pixels. For each run a location of the starting pixel of a run and either its length or the location of its ending pixel must be recorded. Fig. 9 shows runs in a binary fingerprint image.

The following five cases are identified:
Case 1 : There are no adjacent runs both on the previous and the next scan line.
Case 2: There are two adjacent runs on the previous and the next scan line.
Case 3: There is one adjacent run on either the previous or the next scan line.
Case 4: There are two adjacent runs on either the previous or the next scan line.
Case 5: There are more than two adjacent runs on either the previous or the next scan line.

The first case means the run is one pixel spot or an isolated line with more than one pixel width. The second case means the run is part of a ridge flow. The third case means the run is a ridge termination, either the starting or the ending points of ridge flow. The fourth case means two runs on the previous scan line are merging or one run is splitting into two runs on the next scan line. Finally, the fifth case has not been considered in the current experiment because a confluence point which is composed of more than two ridge flows is not a minutia in AFRS. The runs in both the third and the fourth cases are called characteristic runs, whereas the runs in the second case are called regular runs. Characteristic runs of the third case correspond to candidates for termination minutiae in a fingerprint image and those of the fourth case stand for bifurcation in a fingerprint image. Some false minutiae are also detected in the process (fig. 10), hence some post processing is necessary for their validation.

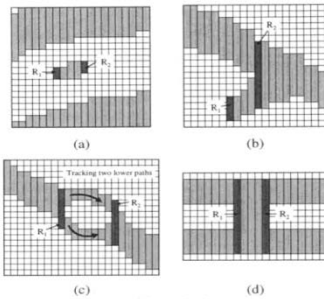


Fig. 10 Examples of false run structures. (a) island (b) spike (c) hole (d) bridge

4.1.1.3 Ridge flow and local pixel analysis

Gamassi et al [49] also proposed a square based method to extract minutiae from unthinned binarized images. Around each pixel in the fingerprint image, the method creates a 3x3 square mask and computes the average of pixels. If the average is lesser than 0.25, the pixel is identified as a ridge termination minutiae and if the average is greater than 0.75, the pixel is treated like a bifurcation minutiae. Alibeigi et al [50] further used this method and proposed a hardware scheme based on pipelined architecture for the same. Maddala et al [51] described the implementation and evaluation of an existing fingerprint recognition system developed by the National Institute of Standards and Technology (NIST). The fingerprints are first enhanced and binarized. The binarized image is then scanned both horizontally and vertically using a 2x3 pixel window size to identify ridge endings and bifurcations. A post processing stage is employed to minimize the number of false minutiae.

4.1.2 Skeletonization-based Minutiae Extraction (Minutiae Extraction from thinned binarized images with Image Post processing)

Here again the image is preprocessed for enhancement. As explained in the previous section the image is segmented and binarized. Next, the binarized image is thinned. The thinning algorithm removes pixels from ridges until the ridges are one pixel wide [93]. There are other methods also available for thinning [94, 95, 96]. Then the minutiae are extracted from the enhanced, binarized and thinned image. Following the extraction of minutiae, a final image post processing stage is performed to eliminate false minutiae. Most of the techniques in this category are based on the concept of crossing number while some are morphology based.

4.1.2.1 Crossing Number

The most commonly employed method of minutiae extraction in this category is the Crossing Number (CN) concept. A large number of techniques for minutiae extraction available in the literature [52-69] belong to this category.

P_4	P_3	P_2
P_5	P	P_1
P_6	P_7	P_8

Fig. 11 3X3 neighbourhood

This method is favored over other methods for its computational efficiency and inherent simplicity. This method involves the use of the skeleton image where the ridge flow pattern is eight-connected. The minutiae are extracted by scanning the local neighbourhood of each ridge pixel in the image using a 3X3 window (fig. 11). The CN value is then computed as follows:

$$CN = 0.5\sum_{i=1}^{8} | P_i - P_{i+1} | \qquad (4)$$

where P_9=P_1. It is defined as half the sum of the differences between pairs of adjacent pixels in the eight-neighbourhood. Using the properties of the CN as shown in fig. 12, the ridge pixel can then be classified as a ridge ending, bifurcation or non-minutiae point. For example, a ridge pixel with a CN of one corresponds to a ridge ending, and a CN of three corresponds to a bifurcation.

CN	Property
0	Isolated point
1	Ridge ending point
2	Continuing ridge point
3	Bifurcation point
4	Crossing point

Fig. 12 Properties of crossing number.

Other authors such as Jain et al. [4] have also performed minutiae extraction using the skeleton image. Their approach involves using a 3X3 window to examine the local neighbourhood of each ridge pixel in the image. A pixel is then classified as a ridge ending if it has only one neighbouring ridge pixel in the window, and classified as a bifurcation if it has three neighbouring ridge pixels. Consequently, it can be seen that this approach is very similar to the Crossing Number method.

False minutiae may be introduced into the image due to factors such as noisy images, and image artifacts created by the thinning process. Hence, after the minutiae are extracted, it is necessary to employ a post processing stage in order to validate the minutiae. Fig. 13 illustrates some examples of false minutiae structures, which include the spur, hole, triangle and spike structures [52]. It can be seen that the spur structure generates false ridge endings; whereas both the hole and triangle structures generate false bifurcations. The spike structure creates a false bifurcation and a false ridge ending point.

The majority of the proposed approaches for image post processing in literature [52, 60, 61, and 62] are based on a series of structural rules used to eliminate spurious minutiae. For example, a ridge ending point that is connected to a bifurcation point, and is below a certain threshold distance is eliminated. However, rather than employing a different set of heuristics each time to eliminate a specific type of false minutiae, some approaches incorporate the validation of different types of minutiae into a single algorithm.

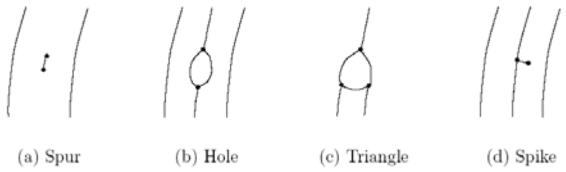

(a) Spur (b) Hole (c) Triangle (d) Spike

Fig. 13 Examples of typical false minutiae structures.

They test the validity of each minutiae point by scanning the skeleton image and examining the local neighbourhood around the minutiae. The algorithm is then able to cancel out false minutiae based on the configuration of the ridge pixels connected to the minutiae point.

In [67], the authors propose fingerprint preprocessing before feature extraction. The preprocessing included obtaining the vertical oriented fingerprint image followed by the core point detection and region of interest selection. Then feature extraction is done in the extracted region of interest image.

Leung et al [68] proposed a neural network based approach to minutiae extraction where preprocessing techniques are first applied to a clean and thinned binary fingerprint ridge structure, which is ready for feature extraction and then a multilayer perceptron network of three layers is trained to extract the minutiae from the thinned fingerprint image.

4.1.2.2 Morphology based

There are minutiae extraction techniques [69, 70] which are based on mathematical morphology. They preprocess the image so as to reduce the effort in the post processing stage. One such technique [70] preprocesses the image with morphological operators to remove spurs, spurious bridges etc. and then uses the morphological Hit or Miss transform to extract true minutiae. Morphological operators are basically shape operators and their composition allows the natural manipulation of shapes for the identification and the composition of objects and object features. The technique develops structuring elements for different types of minutiae present in a fingerprint image to be used by the HMT to extract valid minutiae. Ridge endings are those pixels in an image which have only one neighbour in a 3X3 neighbourhood.

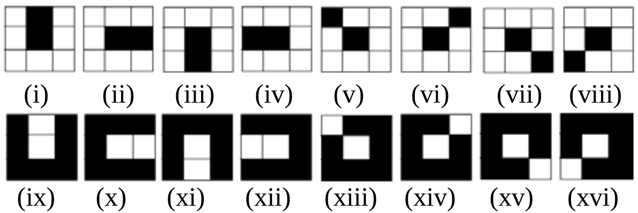


Fig. 14 (i) to (viii) The structuring element sequence J_1 = (J_1^1, J_1^2, J_1^3, J_1^4, J_1^5, J_1^6, J_1^7, J_1^8). (ix) to (xvi) The structuring element sequence J_2 = (J_2^1, J_2^2, J_2^3, J_2^4, J_2^5, J_2^6, J_2^7, J_2^8).

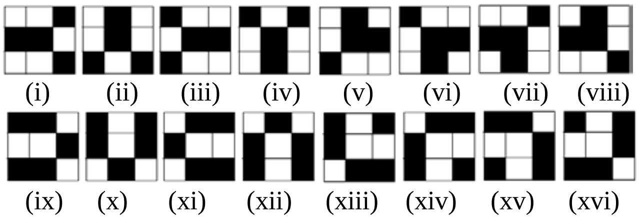


Fig. 15 (i) to (viii) The structuring element sequence $J_1 = (J_1^1, J_1^2, J_1^3, J_1^4, J_1^5, J_1^6, J_1^7, J_1^8)$. (ix) to (xvi) The structuring element sequence $J_2 = (J_2^1, J_2^2, J_2^3, J_2^4, J_2^5, J_2^6, J_2^7, J_2^8)$.

The minutiae image M1 containing ridge terminations is given by applying Hit or Miss transform on I by J as follows:

$$M1 = I \otimes \square\square J \quad (5)$$

where, I is the thinned image and J is the sequence of structuring element pairs (J1, J_2) shown in fig. 14. Ridge bifurcations are those pixels in an image which have only three neighbours in a 3X3 neighbourhood and these neighbours are not adjacent to each other. The minutiae image M2 containing ridge terminations are given by:

$$M2 = I \otimes \square\square J \quad (6)$$

where, I is the thinned image and J is the sequence of structuring element pairs (J_1, J_2) shown in fig. 15. As mentioned earlier, the problem with other techniques is the generation of a large number of spurious minutiae together with true ones whereas this algorithm results in efficient minutiae detection, thereby saving a lot of effort in the post processing stage.

4.2 Minutiae Extraction from Gray-Level images

Minutiae detection can also be done directly from gray-level fingerprint images. A number of techniques exist, but it is still a topic of research. Extracting features directly from a gray scale image without binarization and thinning is of great relevance because of the following reasons:

- A lot of information may be lost during binarization process.
- Binarization and thinning are time consuming.
- The aberrations and irregularity of the binary fingerprint image adversely affect the fingerprint thinning procedure and a relatively large number of spurious minutiae are introduced by the binarization thinning operations.
- Most of the binarization techniques prove to be unsatisfactory when applied to low quality images.

4.2.1 Minutiae Extraction by following ridge flow lines

Based on the observation that a ridge line is composed of a set of pixels with local maxima along one direction, Maio and Maltoni [71, 72] proposed extracting the minutiae directly from the gray-level image by following the ridge flow lines with the aid of the local orientation field. This method attempts to find a local maximum relative to the cross-section orthogonal to the ridge direction. From any starting point $Pt_s(x_c, y_c)$ with local direction θ_c in the fingerprint image, a new candidate point $Pt_{n(}x_n, y_n)$ is obtained by tracing the ridge flow along the θ_c with fixed step of μ pixels from $Pt_s(x_c, y_c)$. A new section Ω containing the point $Pt_{n(}x_n, y_n)$ is orthogonal to $\theta_{c.}$ The gray-level intensity maxima of Ω becomes $Pt_s(x_c, y_c)$ to initiate another tracing step. This procedure is iterated till all the minutiae are found. The optimal value for the tracing step μ and section length σ is chosen based on the average width of ridge lines. Jiang et al. [73] improved the method of Maio and Maltoni by choosing dynamically the tracing step μ according to the change of ridge contrast and bending level. A large step μ is used when the bending level of the local ridge is low and intensity variations along the ridge direction are small. Otherwise a small step μ value is used. Instead of tracing a single ridge, Liu et al. [74] proposed tracking a central ridge and the two surrounding valleys simultaneously. In each cross section Ω a central maximum and two adjacent minima are located at each step, and the ridge following step μ is dynamically determined based on the distance between the lateral minima from the central maximum. Minutiae are extracted where the relation <minimum, maximum, minimum> is changed. Linear Symmetry (LS) filter in [75, 76] is used to extract the minutiae based on the concept that minutiae are local discontinuities of the LS vector field. Two types of symmetries - parabolic symmetry and linear symmetry are adapted to model and locate the points in the gray-scale image where there is lack of symmetry (fig. 16). A window size of 9 × 9 is used to calculate the symmetry filter response. Candidate minutiae points are selected if their responses are above a threshold.

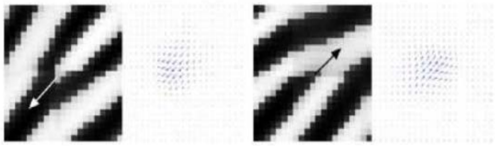


Fig. 16 Symmetry filter response in the minutia point. Left-ridge bifurcation, Right-ridge ending.

Gao et al [77] proposed a minutia extraction method based on Gabor phase. Differing from most existing methods, the approach works in the transform domain of the fingerprint image where, the image is convolved by a Gabor filter, resulting in a complex image. It is then transformed into the amplitude and phase part. A minutiae extractor then extracts minutiae directly from the Gabor phase field. Ratha et al [78] proposed a minutiae extraction algorithm in which the flow direction of ridges is computed by viewing the fingerprint image as a directional textured image. A ridge segmentation algorithm based on a waveform projection is then used to accurately locate the ridges and a thinned ridge image is obtained and smoothed using morphological operators. Finally the minutiae are extracted from the thinned ridges based on the number of crossings and a post processing step applied to remove spurious minutiae.

4.2.2 Fuzzy techniques for minutiae extraction from gray level images

Some fuzzy techniques have also been suggested in literature to extract minutiae from gray scale images directly. Sagar et al [79, 80] proposed that a gray scale image consists of two distinct levels of gray pixels. The darker pixels, constituting the ridges form one such level. The lighter pixels, constituting the valleys and furrows form another such level. Using human linguistics, these two levels of gray can be described as DARK and BRIGHT levels correspondingly. By using fuzzy logic, these two levels are modeled and used along with appropriate fuzzy rules to extract minutiae accurately. For this purpose, rough line thinned structures for both ridges and valleys are obtained. Since bifurcations can be seen as valley endings, the same algorithm that determined ridge endings could be applied to determine valley endings. A 5x5 pixel test window is placed at every point of the line thinned structure. The average value of the 25 pixels is obtained. In addition, the average value of pixels within a 2-pixel border surrounding the test window is also obtained. These two averages form the linguistics variable of brightness. For ridge endings detection, the first average determines the DARK full membership and the second average determines the BRIGHT full membership. This is reversed for the bifurcation detection.

5. Conclusions

Image quality is related directly to the ultimate performance of automatic fingerprint authentication systems. Good quality fingerprint images need only minor preprocessing and enhancement for accurate feature detection algorithm. This paper reviewed a large number of techniques described in the literature to extract minutiae from fingerprint images. The approaches are distinguished on the basis of several factors like: the kind of input images they handle i.e. whether binary or gray scale, techniques of binarization and segmentation involved, whether thinning is required or not and the amount of effort required in the post processing stage, if exists. But low quality fingerprint images need preprocessing to increase contrast, and reduce different types of noises as noisy pixels also generate a lot of spurious minutiae as they also get enhanced during the preprocessing steps. Further, more emphasis is to be laid on defining the local criteria, in order to establish the validity of a minutia point, which is particularly useful during fingerprint matching and adopting more sophisticated identification models, for instance extending minutiae definition by including trifurcations, islands, bridges, spurs etc. Also, the paper leads to the further study of the statistical theory of fingerprint minutiae. In particular approaches can be investigated to determine the number of degrees of freedom within a fingerprint population which will give a sound understanding of the statistical uniqueness of fingerprint minutiae.

References

[1] L. Hong, Y. Wan, and A. K. Jain, “Fingerprint image enhancement: Algorithms and performance evaluation”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20(8), 1988, pp. 777–789.

[2] M. K. Khan, “Fingerprint Biometric-based Self Authentication and Deniable Authentication Schemes for the Electronic World”, IETE Technical Review, Volume 26, Issue 3, 2009.

[3] A. K. Jain, L. Hong, and R. Bolle, “On-line fingerprint verification”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(4), 1997, pp. 302–314.

[4] A. K. Jain, L. Hong, S. Pankanti, and R. Bolle, “ An identity authentication system using fingerprints”. Proc. IEEE, 85(9), 1997, pp.1365–1388.

[5] A. K. Jain, S. Prabhakar, L. Hong, and S. Pankanti, “Filterbank-based fingerprint matching”, Image Processing, IEEE Transactions on, 9(5), 2000, pp. 846–859.

[6] M. Kaur, M. Singh, P.S. Sandhu, “Fingerprint Verification system using Minutiae Verification Technique”, Proceedings of world Academy of Science, Engineering and Technology, vol. 36, 2008.

[7] L. H. Thai and N. H. Tam, “Fingerprint recognition using standardized fingerprint model”, IJSCI International Journal of Computer Science Issues, vol. 7, issue 3, no. 7, 2010, pp. 11-16.

[8] R. Cappelli, A. Lumini, D. Maio, and D. Maltoni, “Fingerprint classification by directional image partitioning”, Pattern Analysis and Machine Intelligence, IEEE Transactions on, 21(5), 2002, pp. 402–421.

[9] R. Cappelli, D. Maio, J. L. Wayman, and A. K. Jain, “Performance evaluation of fingerprint verification systems. IEEE Transactions on Pattern Analysis and Machine Intelligence”, 28(1), 2006, pp. 3–18.

[10] S. Pankanti, S. Prabhakar, and A. K. Jain, “On the individuality of fingerprints”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(8), 2002, pp. 1010–1025.

[11] A. K. Jain, F. Patrick, A. Arun, “Handbook of Biometrics. Springer science and Business media”, I edition, 2008 pp. 1-42.

[12] S. Prabhakar, J. Wang, A. K. Jain, S. Pankanti, and R. Bolle, “Minutiae Verification and classification for fingerprint matching”. Proc. 15th International Conference Pattern Recognition (ICPR) vol. 1, 2000, pp. 25–29.

[13] B. Bir and T. Xuejun, “Fingerprint indexing based on novel features of minutiae triplets”, IEEE Transactions on

Pattern Analysis and Machine Intelligence, 25(5), 2003, pp. 616–622.

[14] Z. Chen and C. H. Kuo, “A topology-based matching algorithm for fingerprint authentication”, in proc. IEEE International Carnahan Conference on Security Technology, 1991, pp. 84–87.

[15] F. Chen, J. Zhou and C. Yang, “Reconstructing Orientation Field from Fingerprint Minutiae to Improve Minutiae Matching Accuracy”, IEEE Transactions on Image Processing, vol. 18, no. 7, 2009, pp. 1665-1670.

[16] K. Cao, X. Yang, X. Tao, P. Li, Y. Zang and J. Tian, “Combining features for distorted fingerprint matching”, Journal of Network and Computer Applications, vol. 33, 2010, pp. 258-267.

[17] H. Choi, K. Choi and J. Kim, “Fingerprint Matching Incorporating Ridge Features With Minutiae”, IEEE Transactions on Information Forensics and Security, vol. 6, no. 2, 2011, pp. 338-345.

[18] S. Kumar D. R., K. B. Raja, R. K. Chhotaray and S. Pattanaik, “DWT Based Fingerprint Recognition using Non Minutiae Features”, IJSCI International Journal of Computer Science Issues, vol. 8, issue 2, no. 7, 2011, pp. 237-264.

[19] N. Goranin and A. Cenys, “Evolutionary Algorithms Application Analysis in Biometric systems”, Journal of Engineering Science and Technology Review vol. 3, no. 1, 2010, pp. 70-79.

[20] T, V. Le, K. Y. Cheung and M. H. Nguyen, “A Fingerprint Recognizer Using Fuzzy Evolutionary Programming”, In Proc. Hawaii International Conference on System Sciences, 2001.

[21] J. Jaam, M. Rebaiaia and A. Hasnah, “A Fingerprint Minutiae Recognition System Based on Genetic Algorithms”, The International Arab Journal of Information Technology, vol. 3, no. 3, 2006, pp. 242-248.

[22] A.A. Paulino, A. K. Jain, F. Jianjiang, "Latent Fingerprint Matching: Fusion of Manually Marked and Derived Minutiae," In Proc. 23rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), 2010 , pp.63-70.

[23] B. G. Sherlock, D. M. Monro, and K. Millard, “Fingerprint enhancement by directional Fourier filtering, Vision”, IEE Proceedings on Image and Signal Processing, 141(2), 1994, pp. 87–94.

[24] Trier, T. Taxt, “Evaluation of binarisation methods for document images”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 17, No. 3, 1995, pp.312–315.

[25] C. Wu, Z. Shi, and V. Govindaraju, “Fingerprint image enhancement method using directional median filter”, Biometric Technology for Human Identification, SPIE, volume 5404, 2004, pp. 66–75.

[26] S. Greenberg, M. Aladjem, D. Kogan and I. Dimitrov, “Fingerprint Image Enhancement using Filtering Techniques”, Real–Time Imaging vol. 8, 2000, pp. 227–236.

[27] Chuen-Horng Lin, W. Lin, “Image Retrieval System Based on Adaptive Color Histogram and Texture Features”, The Computer Journal, 2010.

[28] R. C. Gonzalez and R. E. Woods, “Digital Image Processing”, Prentice Hall, Upper Saddle River, NJ, 2002.

[29] D. Maltoni, D. Maio, A. K. Jain, and S. Prabhakar, “Handbook of Fingerprint Recognition”, 2003 Springer.

[30] S.K. Oh, J.J. Lee, C.H. Park, B.S. Kim, K.H. Park, “New Fingerprint Image Enhancement Using Directional Filter Bank”, Journal of WSCG, vol.11, no.1, 2003.

[31] C. Wu, S. Tulyakov, and V. Govindaraju, “Image quality measures for fingerprint image enhancement”, in Proc. International Workshop on Multimedia Content Representation, Classification and Security(MRCS), volume LNCS 4105, 2006, pp. 215–222.

[32] T. Kamei and M. Mizoguchi, “Image filter design for fingerprint enhancement”, in Proc. International Symposium on Computer Vision, 1995, pp. 109–114.

[33] M. T. Leung, W. E. Engeler, and P. Frank, “Fingerprint image processing using neural networks”, in TENCON 90, IEEE Region 10 Conference on Computer and Communication Systems, vol. 2, 1990, pp. 582–586.

[34] M. Hanmandlu, S. N. Tandon, and A. H. Mir, “A new fuzzy logic based image enhancement”, Biomed. Sci. Instrum, vol. 34, 1997, pp. 590–595.

[35] Y.S. Choi and R. Krishnapuram, “A robust approach to image enhancement based on fuzzy logic’, IEEE Trans. Image Process., vol 6, no. 6, 1997, pp. 808-825.

[36] M. Natchegael , E. E. Kerre, “Fuzzy techniques in image processing”, Springer Verlag, 2000.

[37] A. C. Pais Barreto Marques and A. C. Gay Thome, “A neural network fingerprint segmentation method”, in Proc. Fifth International Conference on Hybrid Intelligent Systems, 2006.

[38] M. T. Yildrim, A. Basturk, “A Detail Preerving type-2 Fuzzy Logic Filter for Impulse Noise Removal from Digital Images”, Fuzzy Systems Conference, FUZZ-IEEE, 2007.

[39] R. Bansal, P. Sehgal, P. Bedi, “A novel framework for enhancing images corrupted by impulse noise using type-II fuzzy sets”, in Proc. IEEE International Conference on Fuzzy Systems and Knowledge Discovery(FSKD’2008) vol. 3, 2008, pp. 266-271.

[40] R. Bansal, Malvika Gaur, Payal Arora, P. Sehgal, P. Bedi, “Fingerprint Image Enhancement Using Type-2 fuzzy sets”, in Proc. IEEE International Conference on Fuzzy Systems and Knowledge Discovery(FSKD’2009), Tianjin, China , vol. 3, 2009, pp, 412-417.

[41] C. Ryu, S. G. Kong and H. Kim, “Enhancement of feature extraction for low-quality fingerprint images using stochastic resonance”, Pattern Recognition Letters, vol. 32, 2011, pp. 107-113.

[42] A. Bazen and S. Gerez, “Segmentation of fingerprint images”, in Proc. Workshop on Circuits Systems and Signal Processing (ProRISC 2001), pp. 276–280.

[43] A. J. Willis and L. Myers, “A cost-effective fingerprint recognition system for use with low-quality prints and damaged fingertips”, Pattern Recognition, vol. 34(2):255–270, 2001.

[44] A.M.Bazen and S.H.Gerez, “Achievement and challenges in fingerprint recognition”, in Biometric Solutions for Authentication i an e-World, 2002, pp.23–57.

[45] L. Coetzee and E. C. Botha, “Fingerprint recognition in low quality images”, Pattern Recognition, vol. 26(10), 1993, pp. 1441–1460.

[46] Zhixin Shi , Venu Govindaraju, “A chaincode based scheme for fingerprint feature extraction”, Pattern Recognition Letters, vol. 27, 2006, pp. 462–468.

[47] Zenzo, L. Cinque, and S. Levialdi, “Run-Based Algorithms for Binary Image Analysis and Processing”, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 18, no. 1, 1996, pp. 83-88.

[48] J Hwan Shin, H. Y. Hwang, S Chien, “Detecting fingerprint minutiae by run length encoding scheme”, Pattern Recognition vol. 39, 2005, pp. 1140-1154.

[49] M. Gamassi, V. Pivri and F. Scotti, “Fingerprint local analysis for high performance minutiae extraction”, IEEE International Conference on Image Processing (ICIP) vol. 3, 2005, pp. 265-268.

[50] E. Alibeigi, M. T. Rizi, P. Behnamfar, "Pipelined minutiae extraction from fingerprint images," Electrical and Computer Engineering, 2009. CCECE '09. Canadian Conference on , vol., no., pp.239-242.

[51] S. Maddala, S. R. Tangellapally, J. S. Bartuněk and M. Nilsson, "Implementation and evaluation of NIST Biometric Image Software for fingerprint recognition," Biosignals and Biorobotics Conference (BRC), 2011 ISSNIP, pp.1-5.

[52] Q. Xiao and H. Raafat, “Fingerprint image postprocessing: a combined statistical and structural approach”, Pattern Recognition vol. 24, no. 10, 1991, pp. 985–992.

[53] J. C. Amengual, A. Juan, J. C. Prez, F. Prat, S. Sez, and J. M. Vilar, “Real-time minutiae extraction in fingerprint images”, in Proc. of the 6th Int. Conf. on Image Processing and its Applications, 1997, pp. 871–875.

[54] A. Farina, Z. M. Kovacs-Vajna, and A. Leone, “Fingerprint minutiae extraction from skeletonized binary images”, Pattern Recognition, vol. 32(5), 1999, pp. 877–889.

[55] J. Xudong and Y. Wei-Yun, “Fingerprint minutiae matching based on the local and global structures”, in Proc. of International Conference on Pattern Recognition (ICPR), vol. 2, 2000, pp. 1038–1041.

[56] M. Tico and P. Kuosmanen, “An algorithm for fingerprint image postprocessing”, in Proceedings of the Thirty-Fourth Asilomar Conference on Signals, Systems and Computers, vol. 2, 2000, pp. 1735–1739.

[57] S. Prabhakar, A. K. Jain, and S. Pankanti, “Learning fingerprint minutiae location and type”, Pattern Recognition, vol. 36(8), 2003, pp. 1847–1857.

[58] S. Shah, P. S. Sastry, “Fingerprint Classification Using a Feedback Based Line Detector”, IEEE Trans. On Systems, Man and Cybernetics, Part B, vol. 34, no.1, 2004.

[59] S. Chikkerur, V. Govindaraju, S. Pankanti, R. Bolle, and N. Ratha, “Novel approaches for minutiae verification in fingerprint images”, in Seventh IEEE Workshops on Application of Computer Vision (WACV/MOTION’05), vol. 1, 2005, pp. 111–116.

[60] Zhao Feng, Xiaou Tang, “Preprocessing and post processing for skeleton-based fingerprint minutiae extraction”, Pattern Recognition vol. 40, 2007, pp. 1270-1281.

[61] F. Zhao and X. Tang, “Preprocessing and postprocessing for skeleton-based fingerprint minutiae extraction”, Pattern Recognition, vol. 40(4), 2007, pp. 1270–1281.

[62] M. Usman Akram , A. Tariq, Shoaib A. Khan, “ Fingerprint image : pre and post processing”, Int. Journal of Biometrics, Vol. 1, No.1, 2008.

[63] B. N. Lavanya, K. B. Raja, K. R. Venugopal, L. M. Patnaik, "Minutiae Extraction in Fingerprint Using Gabor Filter Enhancement," In Proc. International Conference on Advances in Computing, Control, & Telecommunication Technologies, ACT '09, 2009, pp.54-56.

[64] R. Kaur, P. S. Sandhu and A. Kamra, “A Novel Method for Fingerprint Feature Extraction”, In Proc. International Conference on Networking and Information Technology, 2010.

[65] A.R. Patil, M. A. Zaveri, "A Novel Approach for Fingerprint Matching Using Minutiae," In Proc. Fourth Asia International Conference on Mathematical/Analytical Modelling and Computer Simulation (AMS), 2010, pp.317-322.

[66] P. Pathak, “Image Compression algorithms for Fingerprint System”, IJSCI International Journal of Computer Science Issues, vol. 7, issue 3, no. 9, 2010, pp. 45-50.

[67] P. Gnanasivam and S. Muttan, “An efficient algorithm for fingerprint preprocessing and feature extraction”, ICEBT 2010, Procedia computer Science, vol. 2, 2010, pp.133-142.

[68] W. F. Leung, S. H. Leung, W. H. Lau, A. Luk, “Fingerprint Recognition using Neural Networks”, in Proc. IEEE Workshop on Neural Networks for Signal Processing, 1991, pp. 226-235.

[69] V. Humbe, S. S. Gornale, R. Manza and K. V. Kale, “Mathematical Morphology approach for Genuine Fingerprint Feature Extraction”, Int. Journal of Computer Science and Security (IJCSS), vol. 1, 2007, pp. 53-59.

[70] R. Bansal, P. Sehgal, P. Bedi, “Effective Morphological Extraction of True Fingerprint Minutiae based on the Hit or Miss Transform”, International Journal of Biometrics and Bioinformatics(IJBB), vol. 4, 2010, pp. 71-85.

[71] D. Maio and D. Maltoni, “Direct gray-scale minutiae detection in fingerprints”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(1):27–40.

[72] D. Maio and D. Maltoni, “Neural network based minutiae filtering in fingerprints”, in Fourteenth International Conference Pattern Recognition, vol. 2, 1998, pp. 1654–1658.

[73] X. Jiang, W.-Y. Yau, and W. Ser, “Detecting the fingerprint minutiae by adaptive tracing the gray-level ridge”, Pattern Recognition, vol. 34(5), 2001, 999–1013.

[74] L. Jinxiang, H. Zhongyang, and C. Kap Luk, “Direct minutiae extraction from gray-level fingerprint image by relationship examination”, in International Conference on Image Processing(ICIP), vol. 2, 2000, pp. 427–430.

[75] K. Nilsson and J. Bign, “Using linear symmetry features as a pre-processing step for fingerprint images”, in AVBPA, 2001, pp.247–252.

[76] K. K. Hartwig Fronthaler and J. Bigun, “Local feature extraction in fingerprints by complex filtering. In Advances in Biometric Person Authentication”, LNCS, vol. 3781, 2005, pp.77–84.

[77] X. Gao, X. Chen, J. Cao, Z. Deng, C. Liu and J. feng, “A Novel Method Of Fingerprint Minutiae Extraction Based On Gabor Phase”, In Proc. IEEE International Conference on Image Processing, 2010, pp. 3077-3080.

[78] N. K. Ratha, S. Chen, and A. K. Jain, "Adaptive flow orientation-based feature extraction in fingerprint images". Pattern Recognition, vol. 28(11), 1995, pp. 1657–1672.

[79] V. K. Sagar, D. B. L. Ngo, and K. C. K. Foo, "Fuzzy feature selection for fingerprint identification", in proc. 29th Annual International Carnahan.Security Technology, 1995, pp 85-90.

[80] V. K. Sagar and K. J. Beng, "Hybrid fuzzy logic and neural network model for fingerprint minutiae extraction", International Joint Conference on Neural Networks, IJCNN '99., vol. 5, 1999, pp. 3255–3259.

[81] Guang-Ho Cha, "A Context-Aware Similarity Search for a Handwritten Digit Image Database", The Computer Journal, 2009.

[82] A. K. Jain, Y. Chen, and M. Demirkus, "A fingerprint recognition algorithm combining phase-based image matching and feature-based matching", in Proc. of International Conference on Biometrics (ICB), 2005, pp. 316–325.

[83] A. K. Jain, Y. Chen, and M. Demirkus, "Pores and ridges: Fingerprint matching using level 3 features", in Proc. of International Conference on Pattern Recognition (ICPR), vol. 4, 2006, pp. 477–480.

[84] K. Nandakumar and A.K. Jain, "Local Correlation-based Fingerprint Matching", in Proc. ICVGIP, 2004, pp.503-508.

[85] A. Lindoso, L. Entrena, C. López-Ongil, and J. Liu-Jimenez, "Correlation-Based Fingerprint Matching Using FPGAs", in Proc. FPT, 2005, pp.87-94.

[86] D. K. Karna, S. Agarwal, and S. Nikam. "Normalized Cross-Correlation Based Fingerprint Matching". In proc. Fifth International Conference on Computer Graphics, Imaging and Visualisation (CGIV '08), 2008, pp. 229-232.

[87] S. Klein, A. M. Bazen, and R. Veldhuis, "Fingerprint image segmentation based on hidden markov models", in 13th Annual workshop in Circuits, Systems and Signal Processing, 2002.

[88] F. Alonso-Fernandez, J. Fierrez-Aguilar, and J. Ortega-Garcia, "An enhanced gabor filter-based segmentation algorithm for fingerprint recognition systems", in proc. 4th International Symposium on Image and Signal Processing and Analysis(ISPA 2005), pp. 239–244.

[89] X. Chen, J. Tian, J. Cheng, and X. Yang, "Segmentation of fingerprint images using linear classifier", EURASIP Journal on Applied Signal Processing, vol. 4, 2004, pp. 480–494.

[90] E. Zhu, J. Yin, C. Hu, and G. Zhang, "A systematic method for fingerprint ridge orientation estimation and image segmentation", Pattern Recognition, vol. 39(8), 2006, 1452–1472.

[91] Z. Yuheng and X. Qinghan, "An optimized approach for fingerprint binarization", in International Joint Conference on Neural Networks, 2006, pp. 391–395.

[92] N. Otsu, "A threshold selection method from gray level histograms", IEEE Transactions on Systems, Man and Cybernetics, vol. 9, No. 1, 1979, pp. 62-66.

[93] V. Espinos, "Mathematical Morphological approaches for Fingerprint Thinning", in Proc. 36th Annual International Carnahan Conference on Security Technology, 2002, pp. 43-45.

[94] M. Ahmed and R. Ward, "A rotation invariant rule-based thinning algorithm for character recognition", Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 24(12), 2002, pp. 1672–1678.

[95] P. M. Patil, S. R. Suralkar, and F. B. Sheikh, "Rotation invariant thinning algorithm to detect ridge bifurcations for fingerprint identification", in proc. 17th IEEE International Conference on Tools with Artificial Intelligence, 2005.

[96] X. You, B. Fang, V. Y. Y. Tang, and J. Huang, " Multiscale approach for thinning ridges of fingerprint", in Proc. Second Iberian Conference on Pattern Recognition and Image Analysis, volume LNCS 3523, 2005, pp. 505–512.

First Author Roli Bansal is currently pursuing Ph. D. under the joint supervision of Dr. Punam Bedi and Dr. Priti Sehgal from the Dept. of Computer Science, University of Delhi. She is working in the area of fingerprint image enhancement and watermarking. Earlier, she completed her M.C.A. in 1997 and since then she has been working as Assistant Professor in Keshav College, University of Delhi.

Second Author Dr. Priti Sehgal received her Ph.D. in Computer Science from the Department of Computer Science, University of Delhi, India in 2006 and her M. Sc. in Computer Science from DAVV, Indore, India in 1994. She is an Associate Professor in the Department of Computer Science, Keshav Mahavidyalaya, University of Delhi. She has about 17 years of teaching and research experience and has published papers in National/International Journals/Conferences. Dr. Sehgal has been a member of the program committee of the CGIV International Conference and is a life member of Computer Society of India. Her research interests include Computer Graphics, Image Processing, Biometrics, Visualization and Image Retrieval.

Third Author Dr. Punam Bedi received her Ph.D. in Computer Science from the Department of Computer Science, University of Delhi, India in 1999 and her M.Tech. in Computer Science from IIT Delhi, India in 1986. She is an Associate Professor in the Department of Computer Science, University of Delhi. She has about 25 years of teaching and research experience and has published about 110 papers in National/International Journals/Conferences. Dr. Bedi is a member of AAAI, ACM, senior member of IEEE, and life member of Computer Society of India.

Her research interests include Web Intelligence, Soft Computing, Semantic Web, Multi-agent Systems, Intelligent Information Systems, Intelligent Software Engineering, Intelligent User Interfaces, Requirement Engineering, Human Computer Interaction (HCI), Trust, Information Retrieval and Personalization.

Non DTN Geographic Routing Protocols for Vehicular Ad Hoc Networks

Ramin Karimi, Norafida Ithnin ,Shukor Abd Razak , Sara Najafzadeh

Faculty of Computer Science and Information system
University technology Malaysia
Johor, Malaysia

Abstract

Vehicular Ad Hoc Networks are highly mobile wireless ad hoc networks. Routing of data in VANETs is a challenging task due to rapidly changing topology and high speed mobility of vehicles. Geographic routing protocols are becoming popular due to advancement and availability of GPS devices. In this paper, we review the existing non DTN Geographic Routing Protocols for VANETs and also provide a qualitative comparison of them.

Keywords: *Vehicular Ad Hoc Networks, Mobility, Geographic Routing, DTN.*

1. Introduction

Vehicular Ad hoc Networks (VANET), a new technology to build a wireless network between vehicles (V2V) and vehicles to infrastructure(V2I).VANETs are based on short-range wireless communication (e.g., IEEE 802.11) between vehicles[1]. The Federal Communication Commission (FCC) has allocated 75 MHz in 5.9 GHz band for Dedicated Short Range Communication (DSRC). DSRC was conceived to provide architecture for vehicles in Vehicular Network to communicate with each other and with infrastructure. In DSRC, subsequently specialized as Wireless Access in Vehicular Environment (WAVE), GPS-enabled vehicles that are equipped on-board units can communicate with each other. Each vehicle’s wireless network range may be limited to a few hundred meters, so providing end-to-end communication across a larger distance requires message to hop through several nodes.

Routing refers to move a data packet from source to destination and if required the assignment of a path to the destination. In multi-hop regime routing means to forward packets that contain information through other vehicles [14]. This information refers to alerts about events that already happened, like local danger warnings and traffic flow information. If no vehicle is within the communication range a packet is stored and forwarded as soon as a new vehicle comes into reach.

Routing is one of the key research issues in vehicular networks as long as it supports most emerging applications. Recent research showed that existing routing solutions for mobile ad hoc networks (MANETs) are not able to meet the unique requirements of vehicular networks. Thus, a lot of effort has been devoted during the last years to design VANET-specific routing protocols being able to exploit additional information available in VANET nodes [7](e.g., trajectories of nodes, city maps, traffic densities, constrained mobility, etc.).

Geographic routing is a technique to deliver a message to a node in a network over multiple hops by means of position information. Routing decisions are not based on network addresses and routing tables; instead, messages are routed towards a destination location. With knowledge of the neighbors’ location, each node can select the next hop neighbor that is closer to the destination, and thus advance towards the destination in each step.

The rest of the paper is organized as follows. An overview of geographic routing protocols for VANET is presented in section II. A comparison of Non DTN Routing Protocols in VANET will present in section III and a brief overview of security in VANET and Non DTN geographic routing protocols present in section IV and V respectively. Finally this paper is concluded in sectionVI.

2. Overview of protocols

A routing protocol governs the way that two communication entities exchange information; it includes the procedure in establishing a route, decision in forwarding, and action in maintaining the route or recovering from routing failure. This section describes recent *unicast* routing protocols proposed in the literature where a single data packet is transported to the destination node without any duplication due to the overhead concern. Some of these routing protocols have been introduced in MANETs but have been used for comparison purposes or adapted to suit VANETs’ unique characteristics. Because of the plethora of MANET routing protocols and surveys written on them, we will only restrict our attention to MANET routing protocols used in the VANET context. Table-1 illustrates geographic routing protocols in Vehicular Ad Hoc Networks.

In geographic (position-based) routing, the forwarding decision by a node is primarily made based on the position of a packet's destination and the position of the node's one-hop neighbors. The position of the destination is stored in the header of the packet by the source. The position of the node's one-hop neighbors is obtained by the beacons sent periodically with random jitter (to prevent collision). Nodes that are within a node's radio range will become neighbors of the node. Geographic routing assumes each node knows its location, and the sending node knows the receiving node's location by the increasing popularity of Global Position System (GPS) unit from an onboard Navigation System and the recent research on location services [2], respectively. Since geographic routing protocols do not exchange link state information and do not maintain established routes like proactive and reactive topology-based routings do, they are more robust and promising to the highly dynamic environments like VANETs. In other words, route is determined based on the geographic location of neighboring nodes as the packet is forwarded. There is no need of link state exchange nor route setup.

The fundamental principle in the greedy approach is that a node forwards its packet to its neighbor that is closest to the destination. The forwarding strategy can fail if no neighbor is closer to the destination than the node itself. In this case, we say that the packet has reached the *local maximum* at the node since it has made the *maximum* local progress at the current node. The routing protocols in this category have their own recovery strategy to deal with such a failure.

3. Non-DTN Routing Protocols in VANET

The fundamental principle in the greedy approach is that a node forwards its packet to its neighbor that is closest to the destination. The forwarding strategy can fail if no neighbor is closer to the destination than the node itself. In this case, we say that the packet has reached the *local maximum* at the node since it has made the *maximum* local progress at the current node. The routing protocols in this category have their own recovery strategy to deal with such a failure.

3.1 GSR-Geographic Source Routing

Using the location of the destination, the map of the city and the location of the source node, GSR computes a sequence of junctions the packet has to traverse to reach the destination. The protocol aims to calculate the shortest route between origin and destination applying Dijkstra's algorithm over the street map. The calculated path is a list of junctions that the packet should go through [3, 13].

From here, it applies greedy forwarding, where the greedy destination is the position of the next junction of the list. That is, a node forwards the packet to one that is the closest to next junction. Once a junction of the path is reached, the greedy destination is changed to the next junction and greedy forwarding is applied again.

The protocol works in this way until that packet eventually reaches the destination node.

3.2 A-STAR-Anchor-based Street-and Traffic-Aware Routing

This routing follows the approach of anchor-based routing with street awareness. This is having consciousness of the physical environment around the vehicles; the protocol can take wiser routing decisions [4]. On the other hand, the use of anchor-based routing is not novel either. It consists of including within the packet header the list of junctions (anchors) that the packet must traverse. This approach has been employed in the GSR protocol. In fact, A-STAR relies on GSR to perform the routing task.

However, one novelty provided by A-STAR is the inclusion of traffic density information to weigh the streets of the scenario. This contribution modifies the behavior when computing the route of junctions that a packet must go through. In this way, every streets' weights are defined as a function of their traffic density, and the Dijkstra's algorithm is employed to compute the shortest route between source and destination. With this improvement, data packets are expected to be routed through those streets with more vehicles and, therefore, higher connectivity among nodes.

3.3 CAR- Connectivity-Aware Routing

The protocol is aimed at solving the problem of determining connected paths between source and destination nodes. VANETs' nodes present a high degree of mobility, and nodes cannot know the position of the rest of the vehicles due to several well-known scalability problems [5, 13]. This lack of information makes it impossible to determine, a priori, which streets have enough vehicles to allow messages to be routed through them.

CAR's algorithm is designed to deal with these problems, and to do that it is divided into three stages: (i) finding the location of the destination as well as a connected path to reach it from the source node, (ii) using that path to relay messages, and (iii) maintaining the connectivity of the path in spite of the changes in the topology due to the mobility of vehicles.

In the first stage, the source node broadcasts a route request message. The idea behind this initial broadcast is the following. The reception of, at least, one of these route request messages at the destination means that, at least one connected path exists. The destination node answers the route request message with a response message including its current location so that the first problem is solved. But the source node also needs to know the path to reach the destination.

3.4 GPCR- Greedy Perimeter Coordinator Routing

Because nodes are highly mobile in VANETs, node planarization can become a cumbersome, inaccurate, and continuous process. GPCR have observed that urban street map naturally forms a planar graph such that node planarization can be completely eliminated. In this new representation of the planar graph using the underlying roads, nodes would forward as far as they can along roads in both greedy and perimeter mode and stop at junctions where decision about which next road segment to turn into can be determined[6,7].

3.5 GPSR- Greedy Perimeter Stateless Routings

Using this routing is an algorithm that consists of two methods for forwarding packets: *greedy forwarding*, which is used wherever possible, and *perimeter forwarding*, which is used in the regions where greedy forwarding cannot be.

The greedy forwarding algorithm [8] uses packets that carry the locations of their destinations. The packets are stamped by the source node. This way, the packets are always forwarded to the neighbor that is geographically closest to the destination.

The drawbacks of pure greedy forwarding [9]:

- The position accuracy drops if the nodes move (mobility). It is possible that a location server node changes its position and before update process is performed some nodes remain without location server. This may lead to packet loss. Also, due to outdated neighbor table entries excessive re-sending of data may occur.
- Additional network load due to the beacons
- Missing of recovery from failure due to the link-layer broadcast of the beacons.

This leads to failure in transmission, because nodes being close to each other are not recognized as such.

The recovery strategy of the GPSR called *Perimeter Mode* [3,8] is used in order to avoid the lost packets that may occur in pure greedy technique when there is no neighbor available that is closer to the destination than the current forwarding hop. The perimeter mode of GPSR consists of two elements. First, a distributed planarization algorithm that locally transfers the connectivity graph into a planar graph by the removal of “redundant” edges. Second, an online routing algorithm for planar graphs that forwards a packet along the faces of the planar graph towards the destination node.

3.6 CBF-Contention Based Forwarding

Contention-Based Forwarding (CBF) [10] is a mechanism for position-based unicast forwarding, without the use of neighborhood knowledge. Instead, all suitable neighbors of the forwarding node participate in the next hop selection process and the forwarding decision is based on the actual position of the nodes at the time a packet is forwarded. This algorithm eliminates the drawbacks of pure greedy solution.

In position based routing [9] the principle is that the forwarding of the packet, from one hop to another, is done based on the local geographical position of the nodes. Being based on local position information of each node, it is not necessary to create and maintain a global route. Therefore, the algorithm is generally highly scalable and robust against network mobility.

The CBF mechanism [10] uses a contention-based algorithm to determine the next node forwarder and to keep silent the other nodes. Normally, CBF supports unicast routing, but can be used in VANET’s with information dissemination, so that the packet would be disseminated in several directions at the same time. Its main advantages are:

- All relevant nodes are involved in the decision making, i.e.: decision is based on the current position of all neighboring nodes.
- Low overhead, high scalability and high adaptability to the network mobility due to the missing of neighborhood table or knowledge as well as beacons.

3.7 RDGR-Reliable Directional Greedy Routing

RDGR is a reliable position based greedy routing approach which uses the position, speed, direction of motion and link stability of their neighbours to select the most appropriate next forwarding node [11]. It obtains position, speed and direction of its neighbouring nodes from GPS. If neighbour with most forward progress towards destination node has high speed, in comparison with source node or intermediate packet forwarder node, then packet loss probability is increased. In order to improve DGR protocol and increase its reliability, the proposed strategy introduces some new metrics to avoid loss of packets. The packet sender or forwarder node, selects neighbour nodes which have forward progress towards destination node using velocity vector, and checks link stability of those nodes. Finally, it selects one of them which has more link stability and sends packet to it. It uses combination metrics of distance, velocity, direction and link stability to decide about to which neighbour the given packet should be forwarded.

Unlike DGR this approach not only uses the one hop neighbor's position, speed and direction of motion information, it also considers all neighbours position, speed, and direction of motion information and link stability. This routing approach incorporates potential score based strategy, which reduces link breaks, enhances reliability of the route and improves packet delivery ratio.

3.8 LOUVRE-Landmark Overlays For Urban

Lee introduces a routing solution called "Landmark Overlays for Urban Vehicular Routing Environments" (LOUVRE), an approach that efficiently builds a landmark overlay network on top of an urban topology. Also define urban junctions as overlay nodes and create an overlay link if and only if the traffic density of the underlying network guarantees the multi-hop vehicular routing between the two overlay nodes. LOUVRE [7, 12] contains a distributed traffic density estimation scheme which is used to evaluate the existence of an overlay link. Then, efficient routing is performed on the overlay network, guaranteeing a correct delivery of each packet.

4. Security in Vehicular ad hoc network

As in any major public network, VANETs, when deployed without considering the security requirements, lend themselves vulnerable to a host of attacks. The danger involved in possible road accidents and loss of life further impress upon the need for fail-proof security for VANETs. For example, safety-related applications need a high level of security, as a single vehicle sending out false warnings can disrupt the traffic of a whole highway.
A number of research efforts are on in the field of VANET security.

The IEEE Standard 1609.2 specifies security services for the *Wireless Access in Vehicular Environments* (WAVEs) networking stack and for applications that are intended to run over that stack [15, 16]. Services include encryption using another party's public key and non anonymous authentication. The safety-critical nature of many *Dedicated Short-Range Communicatio*ns/WAVE applications makes it vital that services be specified that can be used to protect messages from attacks such as eavesdropping, spoofing, alteration, and replay. It also takes into account the owner's privacy rights. This means the security services must be designed to respect this right and not leak personal, identifying, or linkable information to unauthorized parties.

5. Comparison of Non DTN geographic Routing protocols in VANETs

In table 1 we give a comparison of the existing Geographic Routing protocols in vehicular ad hoc networks. We classified geographic routing protocols based on greedy forwarding. Some protocols are aimed at providing vehicle-to-vehicle services, while others focus on vehicle-to-roadside communication. In the set of characteristics criteria, we categorize based on the various strategies used by each protocol. All of the protocols are position-based, using knowledge of vehicles' positions and velocities to route messages. These protocols also utilize the greedy forwarding strategy for sending messages to the farthest neighbor in the intended direction. We also observe several predictive approaches, where some speculation is made about characteristics of the nodes involved in a route. Some algorithms make predictions on the current locations of nodes based on the last known position, and velocity of the node. Other algorithms use this same information to make predictions about the stability or estimated lifetime of a route. To provide higher rates of delivery in sparse networks, a buffering (carry-and-forward) strategy is often used. In this strategy, a node may hold a packet in a local buffer until a forwarding opportunity is available, instead of simply dropping the packet. We use a similar term, traffic-aware, to refer to a protocol's ability to utilize traffic information to select an efficient route which includes those protocols that make probabilistic assumptions about traffic density by using static knowledge such as bus routes and lane information.
The criterion route-repair or recovery refers to protocols which either uses a strategy to recover from a greedy local optimum in a position-based route or have a mechanism for repairing broken routes. And moving destination is different scenarios and when vehicles can move fast

6. Conclusions

In this paper, we have presented a review of non DTN Geographic Routing Protocol for Vehicular Ad Hoc Networks then we summarized the protocols and categorized them in terms of map required, transport route required, moving destination.

All these protocols utilize the absolute or relative locations of each node to greedily route message toward a next anchor or a destination vehicle.

Acknowledgments

This work was supported by grant from Universiti Teknologi Malaysia,the GUP grant Vote No.Q.J130000.7182.01J52,K-economy.

TABLE 1: QUALITATIVE COMPARISON OF VANET ROUTING PROTOCOLS

characteristics Routing Protocols	Position-based/geographic	Vehicle to Vehicle	Buffering(Carry-and-forward)	Greedy forwarding	Map-required	Transport routes required	Moving destination
GSR	√	√		√	√		
A-STAR	√	√		√	√	√	
CAR	√	√	√	√			
GPCR	√	√		√			
GPSR	√	√		√			
CBF	√	√		√	√		
RDGR	√	√	√	√	√		√
LOUVRE	√	√		√	√		

References

[1] Hirantha Sithira Abeysekera, B., T. Matsuda, and T. Takine, *Dynamic Contention Window Control Mechanism to Achieve Fairness between Uplink and Downlink Flows in IEEE 802.11 Wireless LANs.* Wireless Communications, IEEE Transactions on, 2008. **7**(9): p. 3517-3525.

[2] Flury, R. and R. Wattenhofer. MLS: An efficient location service for mobile Ad Hoc networks. in 7th ACM International Symposium on Mobile Ad Hoc Networking and Computing, MOBIHOC 2006, May 22, 2006 - May 25, 2006. 2006. Florence, Italy: Association for Computing Machinery.

[3] C. Lochert, H. Hartenstein, J. Tian, H. F¨ußler, D. Hermann, and M Mauve, "A routing strategy for vehicular ad hoc networks in city environments," In *Proc. of the IEEE Intelligent Vehicles Symposium 2003*. Columbus, OH, June 2003:156–161.

[4] Seet, B.C., et al., A-STAR: A mobile ad hoc routing strategy for metropolis vehicular communications. Networking 2004, 2004. **3042**: p. 989-999.

[5] Naumov, V. and T.R. Gross. Connectivity-aware routing (CAR) in vehicular ad hoc networks. in IEEE INFOCOM 2007: 26th IEEE International Conference on Computer Communications, May 6, 2007 - May 12, 2007. 2007. Anchorage, AK, United states: Institute of Electrical and Electronics Engineers Inc.

[6] Lochert, C., Mauve, M., F¨ussler, H., and Hartenstein, H., "Geographic routing in city scenarios," SIGMOBILE Mob. Comput. Commun. Rev., vol. 9, no. 1, pp. 69–72, 2005.

[7] Lin, Y.-W., Y.-S. Chen, and S.-L. Lee, *Routing protocols in vehicular Ad Hoc networks: A survey and future perspectives.* Journal of Information Science and Engineering, 2010. **26**(Compendex): p. 913-932.

[8] Karp, B. and H.T. Kung. GPSR: Greedy Perimeter Stateless Routing for wireless networks. in 6th Annual International Conference on Mobile Computing and Networking (MOBICOM 2000), August 6, 2000 - August 11, 2000. 2000. Boston, MA, USA: ACM.

[9] H. Füßler, J. Widmer, M. Mauve, and H. Hartenstein, A novel forwarding paradigm for position-based Routing (with implicit addressing), Germany, 2003.

[10] C.J. Adler, Information dissemination In vehicular ad hoc networks, München, 2006.

[11] K.prasanth, Improved packet forwarding Approach in vehicular ad hoc networks Using RDGR algorithm, 2010.

[12] Lee, K., Le, M., Haerri J., and Gerla, M. (2008), "Louvre: andmark overlays for urban vehicular routing environments," Proceedings of IEEE WiVeC, 2008.

[13] Hassnaa Moustafa , Yan Zhang " Vehicular Networks Techniques, Standards, and Applications " ,CRC press ,2009.

[14] T. Kosch, Technical concept and prerequisites of car-to-car communication, München: BMW

[15] IEEE Vehicular Technology Society, IEEE trial-use standard for wireless access in vehicular environments—security services for applications and management messages, IEEE Std 1609.2™, 2006.

[16] Al-Sakib Khan Pathan " Security ofSelf-Organizing Networks MANET,WSN,WMN,VANET" ,CRC press ,2010.

Ramin Karimi is currently a Ph.D candidate in Department of Computer Science and Information Technology at Universiti Teknologi Malaysia, Johor, Malaysia. He received M.Sc degree in computer engineering from Iran University of Science and Technology in 2006. His research interests include Vehicular Ad Hoc Networks, Mobile ad-hoc networks, security and communication Networks.

Norafida Ithnin is a senior lecturer at Universiti Teknologi Malaysia. She received her B.Sc degree in computer science from Universiti Teknologi Malaysia in 1995, her MSc degree in Information Teknologi from University Kebangsaan Malaysia in 1998 and her PHD degree in computer science from UMIST, Manchester in 2004. Her primary research interests are in security, networks, Mobile ad-hoc networks, Vehicular Ad Hoc Networks.

Shukor Abd Razak is a senior lecturer at Universiti Teknologi Malaysia. His research interests are on the security issues for the Mobile Ad Hoc Networks, Mobile IPv6 networks, Vehicular Ad Hoc Network and network security. He is the author and co-author for many journal and conference proceedings at national and international levels.

Sara Najafzadeh is currently a Ph.D candidate in Department of Computer Science and Information Technology at Universiti Teknologi Malaysia, Johor, Malaysia. She received M.Sc degree in computer engineering from Iran University of Science and Technology in 2006. Her research interests include Vehicular Ad Hoc Networks, Mobile ad-hoc networks and communication Networks.

Developing Adaptive Elearning : An Authoring Tool Design

Said Talhi[1] and Mahieddine Djoudi[2]

[1] Department of Computer Science, University of Batna, Algeria

[2] XLIM-SIC Lab. & IRMA Research Group, University of Poitiers, France

Abstract

Adaptive hypermedia is the answer to the "lost in hyperspace" syndrome, where the user has normally too many links to choose from, and little knowledge about how to proceed and select the most appropriate ones to him/her. Adaptive hypermedia thus offers a selection of links or content most appropriate to the user.
Until very recently, little attention has been given to the complex task of authoring materials for Adaptive Educational Hypermedia. An author faces a multitude of problems when creating a personalized, rich learning experience for each user.
The purpose of this paper is to present an authoring tool for adaptive hypermedia based courses. Designed to satisfy guidelines of accessibility of the W3C recommendation for authors and learners that present disabilities, the authoring tool allows several authors geographically dispersed to produce such courses together. It consists of a shared workspace gathering all tools necessary to the cooperative development task.

Keywords: *Elearning, Adaptive Hypermedia, Accessibility, Cooperative Authoring Systems.*

1. Introduction

One limitation of traditional "static" hypermedia educational applications is that they provide the same page content and the same set of links to all learners.

Due to the differences in background knowledge, learning styles and preferences, individual students may take very different approaches towards learning. Therefore, Adaptive Educational Hypermedia (AEH) have been developed to offer students personalized learning content to improve their learning outcome.

Adaptive educational hypermedia is the answer to the "lost in hyperspace" syndrome, where the learner has normally too many links to choose from, and little knowledge about how to proceed and select the most appropriate ones to him/her.

The domain of AEH is a relatively new direction of research on the crossroads of hypermedia and learner modeling. This domain is an alternative to the traditional "one-size-fits-all" approach in the development of hypermedia systems. Adaptive educational hypermedia systems build a model of the goals, preferences and knowledge of each individual learner, and use this model throughout the interaction with the learner, in order to adapt the hypertext to the needs of that learner. Fig 1 summarizes the Brusilovsky's taxonomy of adaptive hypermedia technologies [1,2].

The year of 1996, the start of the rapid increase in the use of the Word Wide Web, could be considered a turning point in adaptive hypermedia research. The Web, with its clear demand for adaptivity, served to boost adaptive hypermedia research, providing both a challenge and an attractive platform. All the early systems were essentially lab systems, built to explore some new methods, which used adaptivity in an educational context. [1,2].

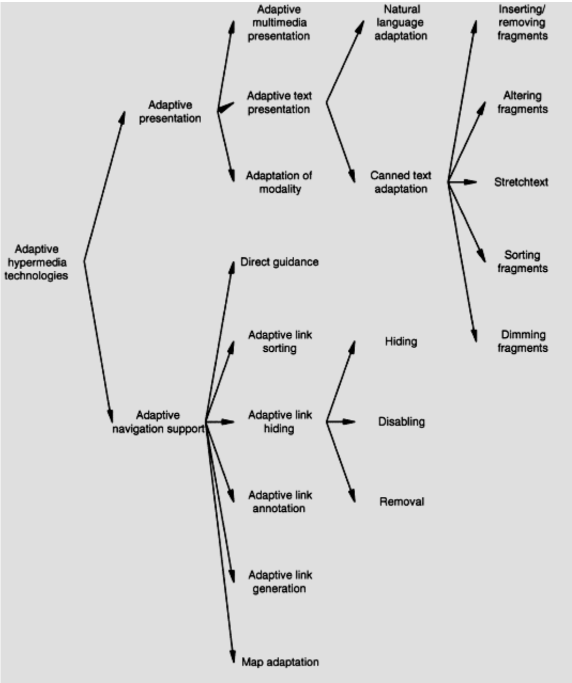


Fig 1. Brusilovsky's taxonomy of adaptive hypermedia technologies.

Despite the efforts carried out during the last few years, the AEH systems development remains a difficult task to undertake. This task requires often the constitution of interdisciplinary group. Experts from different fields such as education and psychology must cooperate with computer engineers to design such systems.

In order to improve the productivity in this domain and allow a wide community to be involved, AEH authoring systems are used and some of them allows the users to develop adaptive hypermedia courses, sometimes, without knowledge in programming art. Thus, the task is reduced in a way that the teachers need only to introduce course material into a generic AEH predetermined by the system.

Authoring tools can enable, encourage, and assist authors in the creation of accessible content through prompts, alerts, checking and repair functions, help files and automated tools. It is just as important that all people be able to author content as it is for all people to have access to it. The tools used to create this information must therefore be accessible themselves.

The authoring tool may be accessible to authors regardless of disability, it produces accessible content by default, and it supports and encourages the author in creating accessible content. Because most of the content of the Web is created using authoring tools, they play a critical role in ensuring the accessibility of the Web. Since the Web is both a means of receiving information and communicating information, it is important that both the Web content produced and the authoring tool itself be accessible

Some of authoring systems are discussed in [3,4], other examples of authoring tools are: [5,6]. However, all these authoring systems were designed to work in a single-user mode.

Recently, thanks to the networks and groupware, virtual meetings involving many people are made possible. Several works in this area are already available in such domains as the cooperative writing [7,8,9], the multimedia, the cooperative design of objects, etc. The common point between all these systems is that they allow several participants to work together in synchronous or asynchronous manner to realize a common task.

Since the cooperative aspect, through a computer network, has been experimented successfully in a lot of domains, this leads us to think that it would be desirable that the designers of authoring tools should integrate this cooperation functionality for AEH production. This is the object of this paper. We investigate this idea through an authoring system called TALABAH (Teaching And LeArning By Adaptive Hypermedia).

We organize the rest of this paper as follows. Section 2 summarizes briefly the concept of accessibility and the authoring systems. Section 3 presents the general concept of authoring systems and discusses the different approaches used when designing cooperative authoring tools; the organizational aspect of our system TALABAH will be presented in this context. Section 4 describes the courseware model and the design of the adaptive hypermedia based course generated by this authoring tool. Section 5 presents the architecture of TALABAH and the different levels and the whole functionalities it covers. Section 6 describes the system implementation and shows some experimental results and discussions. Finally, Section 7 briefly concludes this paper.

2. Accessibility and authoring systems

There is a huge advantage in using authoring tools to create content. In theory, such tools actually promote Web accessibility by allowing easy access to Web content contribution from individuals without expertise in Web authoring.

However, content created by authoring tools can present problems. Often, they do not promote insertion of accessibility features such as alternative text for images. The lack of awareness of many content providers in accessible design issues is accentuated by the relative failure of popular authoring tools to promote the creation of accessible resources.

The W3Cs Authoring Tool Accessibility Guidelines (ATAG) [19] provides a checklist of features with which authoring tools should comply in order to ensure that the Web content they produce is as accessible as possible. A similar effect is noticeable in authoring tools aimed specifically at the learning technology sector, and accessibility of courseware authoring tools is now being addressed.

Even with an authoring tool specifically designed to create fully accessible content, it is vital for content authors to be aware of accessible design techniques, particularly in light of the current constraints affecting Web development environments. Content developers should be aware of the limitations of authoring tools in creating accessible content and should ensure that all resources created are not only designed with accessibility in mind but are checked for accessibility throughout the design lifecycle of the resource [20].

3. From individual authoring to cooperative authoring

The most important shortcoming of an AEH is the authoring part. Developing knowledge space in AEH is not simple and it is very time consuming. In addition, courseware is usually non transferable and non reusable [10]. Some research has been done to address the problems by developing generic authoring systems, for example My Online Tutor [10,11], based on LAOS authoring model [12, 13] that can be delivered by many AEH systems like AHA [14] and WHURLE [10, 15].

Therefore, several works has been taken on the design and implementation of AEH authoring systems during this last decade. Murray [2,3] listed more than twenty references in his state of the art review of the authoring systems dedicated to intelligent tutoring systems and adaptive hypermedia systems. He has classified them in seven categories according to the type of adaptive learning system they produce. These categories are: (1) curriculum sequencing and planning, (2) tutoring strategies, (3) device simulation and equipment training, (4) domain expert system, (5) multiple knowledge types, (6) special purpose and (7) intelligent/adaptive hypermedia.

Given that AEH is often described as having four main components (domain model, adaptive model, learner model, and learner interface), the authoring systems must therefore theoretically include all the necessary tools for building these components. However, it has to be recognized that, very few systems requires from the author to construct every thing needed, the major systems are usually limited to tools for building one, two or at limit three components among the four. The remaining components are generally predefined in a pattern of AEH and the author is solicited only to introduce necessary parameters for their functioning.

TALABAH, the system presented in this paper, generates an adaptive based course that we classify as first and seventh category of the Murray classification mentioned above. This category of authoring systems generally structure the learning material as a network of Learning Units (LUs) where every LU satisfied some educational objectives. The LUs are linked together to show prerequisite-relations between them. Although that these authoring systems do not use any explicit representation of domain knowledge but hypertext representation, they investigated nevertheless the intelligence at the sequencing process of the LUs, the manipulation of the hypertext links and the adaptation of the course according to a student level of knowledge. The LUs to be presented to the learner are then adapted dynamically based on the learner model, the lesson learning objectives and the relations that exist between the different LUs.

On the other hand, given that AEH systems rely generally on large knowledge bases and subject expertise, it would make sense to develop them collaboratively. The models support collaboration works on domain related knowledge for adaptive learning.

To develop a cooperative AEH authoring system, several approaches can be proposed. We can classify them in two large categories [16]. A first approach, pragmatic, and more economic in implementation effort, consists to take an existing single-user authoring system and enrich it with other functionalities that makes it cooperative one. However, the rigidity induced by knowledge acquisition units of the single-user authoring systems, makes it very difficult to take into account group awareness control and the distributed management of the knowledge base. The produced authoring tools will lack certainly effectiveness and will use cooperation mechanisms only at a limited degree.

The second method, which we adopted in the design of TALABAH, consists in taking into account the paradigm of cooperation and the needed tools to do it, at the design step of the system architecture. This approach, although expensive, allows us to apply rigorously the mechanisms of the cooperation metaphor. Though, we must provide through this software architecture, a common work-space to the authors involved in the cooperative construction of an AEH. However, we should notice that the software does not constitute the only aspect in the success of such cooperative system. Also, we have to take into account the human factors involved due to the group activities because of their importance. Thus, to avoid the inherent conflicts due to the human nature, we propose a group organization that allows an optimal way the construction of the AEH.

This organization facilitates also the manipulation of different components of the AEH during all steps of the project advancement. So, we define three roles through which the authors can participate during the AEH building process: main author, constructor coauthor and commentator coauthor.

- The role of the main author is to coordinate the whole work and to verify that the calendar is well respected. He defines the AEH logical structure to be produced by decomposing it in several components (chapters, LU, figures, images, etc.), then he affects the roles to different co-authors. He has free access to all AEH components.

- A constructor coauthor is authorized to create, modify or delete only the components assigned to him. On the remaining AEH components he will have only the role of commentator.

- A commentator coauthor is authorized only to read and /or comment the components assigned to him.

3. The adaptive hypermedia based course

3.1. Courseware model

Two learning modes are presented to the learner in the adaptive learning environment (ALE): “information mode” (free exploration) and “training mode” (learning with auto-evaluation). The learning process is organized around adaptive hypermedia components. The learning material is structured in three abstraction level hierarchy according to three level hierarchy of learning objectives defined in [17]: parts (satisfying the general objectives), chapters (satisfying the specific objectives) and the Hypermedia Learning Units (HLU) (satisfying the operational objectives).

To intelligently sequencing the curriculum and adapt it to each learner capacities, the management of these components, is ensured by rule based system that use five sets of production rules. These rules (for which parameters can be set), called “Main Rules” (MRules), describe the different tutoring plans depending on the different learning situations. They constitute therefore a generic knowledge base that is instantiated in a suitable way for each AEH created by TALABAH.

The instantiation process, producing “Generated Rules” (GRules), is carrying out automatically by the system on the base of parameters delivered by authors. These AEH parameters which are represented in predicates form, describe the quantitative aspect of the teaching material (number of parts, number of chapters, number of learning units, number of questions, number of exercises, etc.).

For reusability and independence from the domains criteria, the Main Rules invoke abstracted structures called Hypermedia Learning Units (HLUs). These HLUs have no knowledge about the AEH domain. They are supposed to receive all kinds of knowledge about the domain via instantiation, under all media types that are allowed by the X/HTML language (text, image, sound, video, applet).

To summarize, we can consider, two levels of knowledge in the curriculum definition:

- **Level 1.** A higher level corresponding to the tutoring plans: These plans consist of five sets of rules that invoke HLU of the lower level. Every set of rules has a specific function. These functions are the following: “negotiation” of the start entry point in the course and/or the objectives to reach; “deduction” of HLUs assumed to be understood after a negotiation phase; “planning” the learning session; “searching and filtering” the content of HLU; and finally “auto-evaluation”.

- **Level 2.** A lower level corresponding to the HLU space: This space consists of a hierarchical network that is constituted of six HLU sub-levels where the first four sub-levels correspond to the courseware-type HLU (module abstract, part abstract, chapter abstract, HLU classes) and the last two sub-levels correspond to evaluation-type HLU (questions and exercises).

3.2. Adaptive learning environment architecture

The adaptive learning environment (ALE) architecture is composed of five (5) modules:

1. A “free-exploration module” that allows the learner to navigate freely through the different HLUs, as a book.

2. Three modules representing the “training mode”:
   - A “domain-expert module” using generated rules to search and filter concept-indexed HLU asked by a pedagogical module at a given moment.
   - A “pedagogical module” that allows the negotiation of learning session objectives with the learner and generates in turn sequencing plan for the adaptive presentation of the lesson. Two sub-modules realize these two tasks: the “negotiator” using the negotiation generated rules and the “planer” using the planning generated rules.
   - A “diagnosis module” that allows the learner evaluation and the maintenance of an overlay type learner model. This module is made up of three sub-modules: an “evaluator” using evaluation generated rules; a “deduction agent” using the deduction generated rules; and a “learner model manager” managing its persistent content.

3. A “supervisor module” that allows on one side, the communication with the learner, and on another side, the coordination between the three modules: domain-expert module, pedagogical module and diagnosis-module. This coordination is carried out via message sending.

4. The cooperative authoring of the course

In order to motivate more authors to use the adaptive hypermedia, the authoring process should be made much simpler than in some existing GUI-based authoring tools. The authoring component should enable the straightforward creation of concepts, the linkage of concepts by prerequisite relationships, and easy generation of the test questions. It should be user-friendly enough to enable a person who is not a computer expert to design the courseware. That includes the development of a graphic editor for concept networks, which will enable the authors to define the prerequisite relationships with a drag-and-drop interface.

From an author point of view, building a courseware using TALABAH consists in introducing, via a cooperative editor, a set of objects that will be manipulated in the adaptive learner environment (ALE). These objects are made up with learning material in the form of hypermedia learning units (HLU), prerequisite-network in the form of an oriented graph, course parameters in the form of predicates and pedagogical knowledge in the form of production rules.

The cooperation task in TALABAH is introduced at the editing level of the teaching material and at the editing level of the prerequisite-network. These two components are well structured: the teaching material is organized as parts, chapters and HLU, and the prerequisite-network is organized as sub-networks form (part-prerequisite network, chapter-prerequisite network, HLU-prerequisite-network and concept-prerequisite network).

These structures are well convenient for the fragmentation and then constitute the basis of our cooperative editing approach as in Alliance [7]. The two concept-keys on which is based the design of TALABAH are the “fragmentation” and “edition roles” [7]. As previously said, we defined three edition roles of participation for the authors: main author, constructor coauthor and commentator coauthor.

At the beginning of the course construction task, a negotiation step is necessary. The main author assigned the edition roles to different co-authors around different fragments of course structure in accordance with their competences and availability. Five learning principles had been incorporated into the authoring process [18]: a clear definition of educational objectives, definition of pre-requisite knowledge, providing a variety of presentation styles (tell, show and do), enhanced feedback and testing, and permitting the learner to control the direction of the learning session by choosing himself the educational objectives.

4.1. Cooperation modes and group awareness

The cooperative developing process of the course is characterized by a steps-sequence during which the authors can work either individually or collectively. In this way, we defined three cooperation modes: individual responsibility, alternate version and collective responsibility. The first two modes are typically asynchronous cooperation modes. Especially, the second one is inspired from the real principle “let us reflect separately on the question and then compare our results after”.

The last one is a typically synchronous mode that allows, to relatively reduced number of authors chosen by main author, to finalize the course version when the project reaches its final phase [21]. The notification and group awareness functions constitute an important point in the cooperative application design [22]. It includes all the interface functions and all systems functions that allow the users to perceive the activities of the other users, as well as to control and to act on the distributed environment.

4.2. The cooperative editor architecture

The cooperative editor is organized according to centralized client/server architecture where all the communications pass automatically through the central site (the server). We associate to every client-site a client process (CPR) that accomplishes all the tasks that are processed locally (the editing tasks for example). We define a server process (SPR) that manages all the communications between the different CPRs and keeps up to date the content of the course central copy and the course logical structure.

The software architecture offers several functionalities that we can decompose them in three layers: server layer, editor layer and presentation layer. Every layer is structured as a collection of modules where each module consists of several objects implementing some functionalities (see Figure 2 for client side and Figure 3 for server side).

The need for information exchange between the two client layers on one side, and between the client and the server on the other side, implies the presence for “dialog controllers”. We interpose therefore between every presentation layer and every editor layer a Dialog Controller (DC), and between the server layer and every editor layer a Main Dialog Controller (MDC). Messages exchanged between layers transits automatically by the dialog controllers.

According to message-type, the convenient objects are then executed among those that are defined in a layer.

1 - Presentation Layer (Figure 2): This layer gathers an organized set of interactive objects defining the graphical user interface (buttons, icons, scrolling bar, pull-down menus, etc.). Thus, for every object, modeling a part of our application domain, we associate a presentation technique accomplished by a reactive object that reacts to the different authoring actions.

Besides the pull-down menus achieving the different functions, we especially find a toolbox containing graphical icons that refer to the frequently used functions and specialized widget-based palette allowing the graphical construction of the prerequisite-networks.

2 - Editor Layer (Figure 2): This software layer gathers many types of functionalities allowing every author to manipulate the objects that constitutes the course. These functionalities include not only the support of individual actions, but also the sharing aspect and transparency management. For example, the access to a file in a single user editor delivers directly its content. But in our case, this process consists in several tasks such as access rights verification, locking state of the object and warning the authors working on this file in the same time.

At each author site, some associated functionalities allow the author to save locally the objects that are accessible to him. He will solicit regularly the server to update the versions of these objects. The components of the editor layer are:

a) **HLU/prerequisite-network-Editor:** Two modules are designed to implement this component software. They allow the creation task and the maintenance of different course objects. The first module allows the wisiwig HLU edition using X/HTML language. The second module allows the author to edit the prerequisite-network in a graphical form. This oriented network is made up of linked nodes where the links indicates the different possible progressions between the teaching material components. Four levels are used in the network. One level shows the concept-prerequisites, the second shows parts-prerequisites, the third one shows the chapter-prerequisites of a particular part, and the fourth level shows the HLU-prerequisites of a chapter.

b) **Parameters acquisition module:** This module allows the main author to specify the course parameters that indicate the manner in which the teaching material is decomposed (number of parts, number of chapters in every part, number of HLU in every chapter, etc.). These parameters are saved in the predicates form and then used to instantiate the Main Rules. For example the predicate nbhlu(1,2,4) indicates that chapter 2 in part 1 contains 4 HLUs.

c) **GRules generator module:** This module allows the author to generate the five packages of generated rules that represent different tutoring plans. Based on the course parameters introduced via the previous module, this generation consists of an instantiation of the five packages of the MRules.

d) **Verification module:** As most authoring systems, TALABAH offers a tool to help the author in the diagnosis of errors and bugs. It facilitates detection of incoherencies that can be occurred during the course construction. For example, at the end of the construction process of the course, it is necessary to check the compatibility of course parameters with the effective structure of the teaching material.

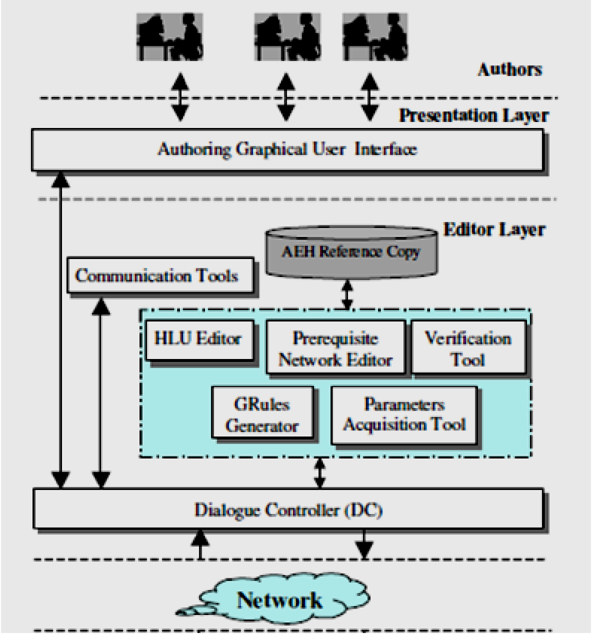


Fig 2. Architecture of the system - Client Side

3 - Server Layer (Figure 3): This software layer gathers several types of functionalities, among which those that concern the course logical structure management, as well as the content of the course components. They allow the authors to save and retrieve course objects whose logical structure is declared, as well, at the central level as at the local level. This software layer is responsible for access rights control, events handling and events notification. In the case of events notification, for example, the concerned module manages a set of queues such as engagement queue, locking-queue, etc. At every time, if an event occurs, this process identifies the concerned authors and proceeds to structure the notifications as a message form to transmit. These messages then will be made available to another sender module that sends the message.

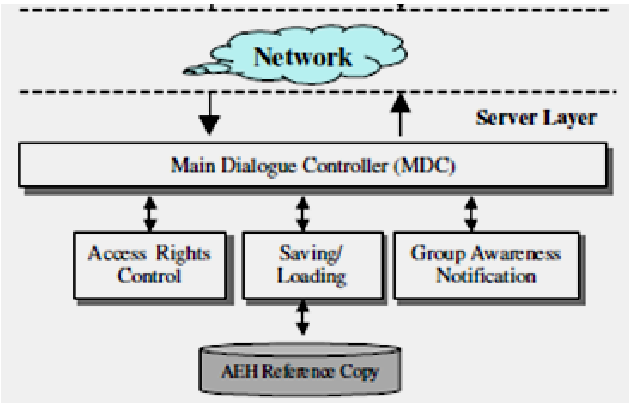


Fig 3. Architecture of the system - Server Side

4 - Dialogue Controllers DC and MDC (Figure 4): Each dialog controller is composed of three independent modules performing respectively, "message reception", "message control" and "message sending". The Control module allows the coordination and synchronization of the running of the different modules within the three layers, in accordance with the actions of the different authors. At any time, it used all necessary information to determine exactly what are the functionalities to invoke within the layers for which it is responsible. Every time that an event occurs, the associated receiver delivers the message materializing this event to the control module. The control module reacts then following three steps: analyze the event, draw up an action plan and then carry out the established plan.

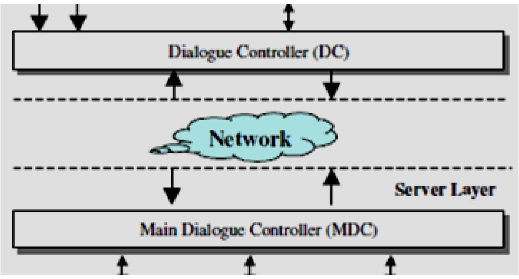
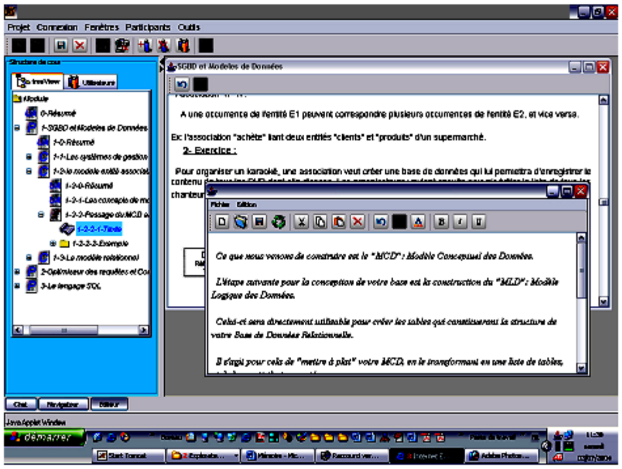


Fig 4. Dialogue Controllers between server and client layers

5. Discussion

The AEH (course) pattern is implemented in PHP/MySQL and resides on a server; it can therefore be accessed simultaneously by different distant learners. The authoring tool, implemented in JSP and Java, is organized as centralized client-server architecture (Figure 5 shows an interface screen). It makes possible to several authors to be connected to a working session characterized by a cooperative space and a control strategy. The cooperation space is represented by a set of structured components (HLU, prerequisite-networks, course parameters and the five packed rules) and tools, which make it possible the edition and communication tasks.



A une occurrence de l'entité E1 peuvent correspondre plusieurs occurrences de l'entité E2, et vice versa.

Ex: l'association "achète" liant deux entités "clients" et "produits" d'un supermarché.

2- Exercice :

Pour organiser un karaoké, une association veut créer une base de données qui lui permettra d'enregistrer le contenu

chanteur

Ce que nous venons de construire est le "MCD" : Modèle Conceptuel des Données.

L'étape suivante pour la conception de votre base est la construction du "MLD" : Modèle Logique des Données.

Celui-ci sera directement utilisable pour créer les tables qui constitueront la structure de votre Base de Données Relationnelle.

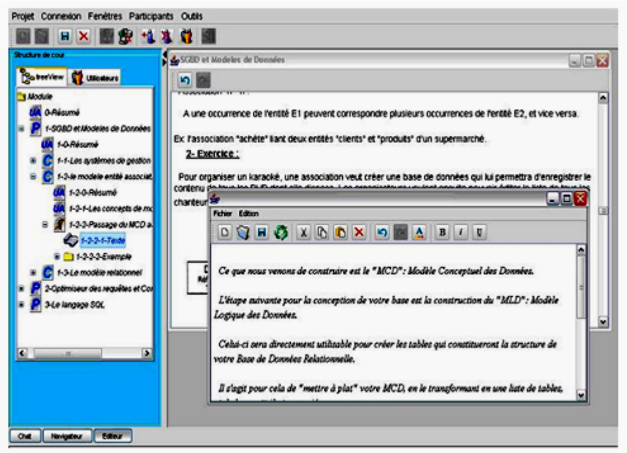
Il s'agit pour cela de "mettre à plat" votre MCD, en le transformant en une liste de tables,

Fig 5. An interface-screen of the authoring tool

The control strategy manages the negotiation of the access right to a component of the AEH and then participation of users during the work session. Five learning principles had been incorporated into the authoring process [19]. These principles are: a clear definition of educational objectives, definition of pre-requisite knowledge, providing a variety of presentation styles (tell, show and do), enhanced feedback and testing, and permitting the learner to control the direction of the learning session by choosing himself the learning objectives.

Two different approaches were used to test the validity that the system actually incorporated pedagogy and effective cooperative design concepts as part of the developmental process. To evaluate the system, a group of four teachers were surveyed to seek their opinion if the authoring system did incorporate the five learning principles into its design. Their survey results validated that the system would prompt developers to build a course based on pedagogy. In addition a high agreement was noted in the self-direction of the lesson offered to the learner.

In a second means to validate the system, five teachers geographically dispersed were invited to develop a course on the "Relational Data Bases", via local network, and were surveyed to seek their opinion if the authoring system offers all cooperative tools necessary to construct the course in a synchronized manner (Figure 6). We were also interested in the group interaction through accounting of various exchanges operated between the authors during a work session. Especially, we record the aspects related to notification and group awareness.


Projet Connexion Fenêtres Participants Outils
Structure de cour
treeView
Utilisateurs
Module
0-Résumé
1-SGBD et Modeles de Données
1-0-Résumé
1-1-Les systèmes de gestion
1-2-le modele entité associat
1-2-0-Résumé
1-2-1-Les concepts de mo
1-2-2-Passage du MCD a
1-2-2-1-Texte
1-2-2-2-Exemple
1-3-Le modèle relationnel
2-Optimiseur des requêtes et Co
3-Le langage SQL
SGBD et Modeles de Données
Fichier Edition
Chat Navigateur Editeur

A une occurrence de l'entité E1 peuvent correspondre plusieurs occurrences de l'entité E2, et vice versa.

Ex: l'association "achète" liant deux entités "clients" et "produits" d'un supermarché.

2- Exercice :

Pour organiser un karaoké, une association veut créer une base de données qui lui permettra d'enregistrer le contenu

chanteur

Ce que nous venons de construire est le "MCD" : Modèle Conceptuel des Données.

L'étape suivante pour la conception de votre base est la construction du "MLD" : Modèle Logique des Données.

Celui-ci sera directement utilisable pour créer les tables qui constitueront la structure de votre Base de Données Relationnelle.

Il s'agit pour cela de "mettre à plat" votre MCD, en le transformant en une liste de tables,

Fig 6. One “Data bases” course developing screen

Although the system does exhibit positive results after a pilot test in the local network context, a question for future research is the experimentation of the system in the internet/web context. This research would provide evidence that the concepts incorporated into the system do impact learning in a positive manner. On the positive side the survey results from the two different experimentations provides indication that the system is a positive benefit to teachers and developers of adaptive educational hypermedia.

6. Conclusion

The legislation in Algerian universities was introduced to ensure that disabled people have the same opportunities as non-disabled people and it is expected that the educational community should do as much as possible to ensure that this happens. Assistive technologies have an important role to play in ensuring that inclusive learning is available to all students without discrimination.

In this way, we have presented an authoring tool that assist disabled users to access teaching and learning activities over the web. The cooperative authoring system, called TALABAH, is designed to satisfy guidelines of accessibility of the W3C recommendation for disabled authors and learners especially with mobility impairments. Integrating cooperation paradigm in AEH authoring systems is the original idea of this paper. This authoring tool allows geographically distant disabled authors to cooperate to produce an accessible courseware according to a predefined course pattern.

Acknowledgments

The work reported in this paper is supported by the CNEPRU project funded by the Algerian Ministry of Higher Education and Scientific Research via contract no. B*01320080006.

References

[1] P. Brusilovsky, “Adaptive hypermedia. methods and techniques of adaptive hypermedia”. International Journal of User Modeling and User-Adapted Interaction, 11 (1/2), 2001, pp. 87-110.

[2] P. Brusilovsky, Developing adaptive educational hypermedia systems: From design models to authoring tools. In Murray, T., Blessing, S., Ainsworth, S. (Eds.), Authoring Tools for Advanced Technology Learning Environment, Dordrecht: Kluwer Academic Publishers, 2003 , pp. 377-409.

[3] T. Murray, “Authoring knowledge-based Tutors: Tools for content, instructional strategy, student model and interface design”, Journal of the Learning Sciences, vol. 7, n° 1, 1998.

[4] T. Murray, “Authoring Intelligent Tutoring Systems: an analysis of the state of the art”, International Journal of AI in Education, vol. 10, 1999, pp. 98-129.

[5] A. Cristea, “Adaptive Course Authoring: My Online Teacher”, in Faculty Mathematics and Computer Science, TU Eindhoven, 2003.

[6] P. De Bra, “AHA! The Adaptive Hypermedia Architecture”, in the ACM Hypertext Conference, Nottingham, UK, 2006.

[7] D. Decouchant, and A.M. Martínez, “A Cooperative, Deductive and Self-Adaptive Web Authoring Environment”, In Proceedings of Mexican International Conference on Artificial Intelligence (MICAI-2000), Springer Verlag, Acapulco (Mexico), 2000, pp. 443-457.

[8] F. Pacull, A. Sandoz, and A. Schiper, “Duplex: a distributed collaborative editing environment in large scale”, Proceedings of the ACM Conference CSCW’94, ACM Press, 1994.

[9] A. Zidani, M. Boufaida, and M. Djoudi, “JamEdit: un outil interactif et coopératif pour l'édition coopérative de documents”, Revue Technique et Science Informatiques, Vol. 19, n°1, Hermes, 2000, pp. 1-23.

[10] C. Stewart, A. Cristea, and T. J. Brailsford, “Authoring Once, Delivering Many: Creating Reusable Adaptive Courseware”, in 4th IAESTED International Conference on Web Based Education (WBE'05), Grindewald, Switzerland,. 2005.

[11] F. Ghali, “Collaborative Adaptation Authoring and Social Annotation in MOT”, Warwick Postgraduate Colloquium in Computer Science, 2008.

[12]. A. Cristea, and A.D. Mooij, “LAOS: Layered WWW AHS Authoring Model and their corresponding Algebraic Operators”, in WWW 2003. ACM Budapest, Hungary, 2003.

[13] N. Weibel, “Towards Adaptive Hypermedia Authoring from the Dexter Model to Laos”, Institute for Information Systems, ETH Zurich., 2006.

[14] A. Cristea, D. Smits, and P. de Bra, “Writing MOT, Reading AHA! - converting between an authoring and a delivery system for adaptive educational hypermedia”, Faculty of

Mathematics and Computer Science, Eindhoven University of Technology, 2005.

[15] M. Meccawy, "WHURLE 2.0: Adaptive Learning Meets Web 2.0", in Third European Conference on Technology Enhanced Learning, Springer: Maastricht, The Netherlands, 2008.

[16] S. Talhi, M. Djoudi, and A. Zidani, "Un système auteur de tuteurs intelligents : évolution du mono-usager vers la coopération", Revue Techniques et Sciences Éducatives, Volume 8, n° 1-2, 2001, pp. 127-138.

[17] D. Hameline, Les objectifs pédagogiques en formation initiale et en formation continue, Edition ESF, 8ième édition, Paris, 1990.

[18] T. Janicki, and Jens O. Liegle, "Development and evaluation of a framework for creating web-based learning modules: a pedagogical and systems perspective", JALN Journal, Volume 5, Issue 1. 2001.

[19] W3C, "Authoring Tool Accessibility", Guidelines, available from: www.w3.org/TR/ATAG10., 2000.

[20] J. M. Slatin, and S. Rush, Maximum Accessibility: Making Your Web Site More Usable for Everyone, Addison Wesley Editor, 2002.

[21] S. Talhi, M. Djoudi, and M. Batouche, "Authoring Groupware For Intelligent Tutoring Systems", Information Technology Journal (ITJ), vol. 5, n°. 5, 2006, pp. 860-867.

[22] A. Muhammad, A. M. Martínez, and D. Decouchant., "Awareness and Coordination for Web Cooperative Authoring". In Proceedings of AWIC'2005, The 3rd International Atlantic Web Intelligence Conference, Lecture Notes in Artificial Intelligence, no 3528, Springer Verlag, Lodz, Poland., 2005, pp.327-333.

Said Talhi received a PhD in Computer Science from the University of Batna, Algeria, in 2007 and he is currently an Associate Professor at the University of Batna, Algeria. His research interests began with intelligent tutoring systems and knowledge based systems in 1992. As Information and Communication Technology became an integral component of any successful education process, his research interests focused on elearning/distance education, authoring systems, collaborative learning and adaptive hypermedia systems where some papers are published. He also published a book in these research fields at EUE (Editions Universitaires Europennes) publisher in 2011.

Mahieddine Djoudi received a PhD in Computer Science from the University of Nancy, France, in 1991. He is currently an Associate Professor at the University of Poitiers, France. He is a member of SIC (Signal, Images and Communications) Research laboratory. He is also a member of IRMA E-learning research group. His PhD thesis research was in Arabic Continuous Speech Recognition. His current research interests is in E-Learning, Web mining and Information Literacy. His teaching interests include Laboratory Information Management System, Data Bases, Quality Management, Web Technology and Computerized System Validation. He started and is involved in many research projects which include many researchers from different Algerian Universities.

Fuzzy-Genetic Classifier algorithm for bank's customers

Elawady R.M.[1], Asim S.A. [2], and Sweidan S.M.[3]

[1]Department of Communication, Faculty of engineering, Mansoura University,
Mansoura, Egypt

[2]Department of information system, faculty of computers &information system,
Mansoura University, Mansoura, Egypt

[3]Department of information system, faculty of computers &information system,
Mansoura University,

Abstract

Modern finical banks are running in complex and dynamic environment which may bring high uncertainty and risk to them. So the ability to intelligently collect, mange, and analyze information about customers is a key source of competitive advantage for an E-business. But the data base for any bank is too large, complex and incomprehensible to determine if the customer risk or default. This paper presents a new algorithm for extracting accurate and comprehensible rules from database via fuzzy genetic classifier by two methodologies fuzzy system and genetic algorithms in one algorithm. Proposed evolved system exhibits two important characteristics; first, each rule is obtained through an efficient genetic rule extraction method which adapts the parameters of the fuzzy sets in the premise space and determines the required features of the rule, further improve the interpretability of the obtained model. Second, evolve the obtained rule base through genetic algorithm. The cooperation system increases the classification performance and reach to max classification ratio in the earlier generations.

Keywords: *fuzzy system, genetic algorithm, rule extraction, E-business*

1. Introduction

Bank customer classification plays an important role for commercial banks to keep away from default risks in customer loan market. A complete customer profile has two parts; factual and behavioral. The factual profile contains information such as name, gender, date of birth that personalization system obtained from the customer's factual data. The behavioral profile models the customer's actions and is usually derived from transactional data. Personalization begins with collecting customer data from various sources, web purchasing, and browsing activities [1]. After the data is collected, it must be stored in the data warehouse. Extracting rule from a given database for cluster customers data is important, there are several algorithms proposed by several researchers,[2,7] used computational intelligence, neural network, genetic algorithms, swarm intelligence (PSO), fuzzy system, rough sets for extracting accurate rules that solve the problem of classification customers with large and incomprehensible database.[5] employed BP neural network to classify customers into 5 groups according to the actual need, [6]use genetic algorithm to predict customer purchasing behavior.

In practical customer classification, there are two problems that can influence the accuracy. On the one hand, in credit scoring areas, we usually cannot label one customer as absolutely good who is sure to repay in time, or absolutely bad who will default certainly. On the other hand, there usually exist many irrelevant variables in the sample data. These redundant irrelevant variables spoil the classification, and increase many unwanted calculations and decrease the accuracy of customer classification. A good computerized classification tool should possess two characteristics, which are often in conflict. First, the tool must attain the highest possible performance, i.e. classify the presented cases correctly as being either *normal* or *risk*. Moreover, it would be highly desirable to be in possession of a so-called *degree of confidence*: the system not only provides a binary classification (*normal* or *risk*), but also outputs a numeric value that represents the degree to which the system is confident about its response. Second, it would be highly beneficial for such a classification system to be human-friendly, exhibiting so-called *interpretability*. The proposed method combines two methodologies fuzzy systems and genetic algorithms which exhibit two important characteristics; first, each rule is obtained through an efficient genetic rule extraction method which adapts the parameters of the fuzzy sets in the premise space and determines the required features of the rule, further improve the interpretability of the obtained model. Second, evolve the obtained rule base through genetic algorithm by enabling the automatic production of fuzzy systems based on a database of training cases (fitness) for extract good rules. The cooperation among the fuzzy system and genetic algorithm increase the classification performance and reach to max classification ratio in the earlier generations.

2. Preliminaries

2.1 Fuzzy systems

Fuzzy logic is a computational paradigm that provides a mathematical tool for representing and manipulating information in a way that resembles human communication and reasoning processes. A fuzzy variables (also called a *linguistic variable*) is characterized by its name tag, a set of *fuzzy values* (also known as *linguistic values* or *labels*), and the membership functions of these labels; these latter assign a membership value, µ label (*u*) to a given real value *u*(*R*, within some predefined range known as the universe of discourse). While the traditional definitions of Boolean logic operations do not hold, new ones can be defined. Three basic operations, and, or, and not, are defined in fuzzy logic as following equations:

$$\mu A\ (u) and\, \mu B\ (u) = min\ \{\mu A\ (u), \mu B\ (u)\ \} \qquad (1)$$

$$\mu A\ (u)\ \ or\ \ \mu B\ (u) = max\ \{\mu A\ (u), \mu B\ (u)\ \} \qquad (2)$$

$$\mu not\ A(u)\ =\ \neg\mu A(u)\ =\ 1 - \mu A(u) \qquad (3)$$

Where *A* and *B* are fuzzy variables. Using such fuzzy operator's one can combine fuzzy variables to form fuzzy-logic expressions, in a Boolean logic. For example, in the domain of control, where fuzzy logic has been applied extensively, one can find expressions such as: **if** room temperature **is** Low, **then** increase ventilation fan speed [8, 9, 10, and 15]. A *fuzzy inference system* is a rule-based system that uses fuzzy logic, rather than Boolean logic, to reason about data. Its basic structure includes four main components, as depicted in Figure ("1"): (1) a fuzzifier which translates crisp (real-valued) inputs into fuzzy values; (2) an inference engine that applies a fuzzy reasoning mechanism to obtain a fuzzy output; (3) a defuzzifier which translates this latter output into a crisp value; and (4) a knowledge base which contains both the rule base, and the initial database. The decision-making process is performed by the inference engine using the rules contained in the rule base. These fuzzy rules define the connection between input and output fuzzy variables [7].

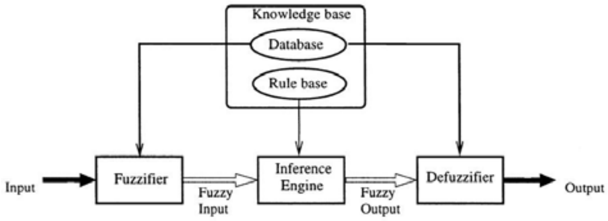


Fig.1 Basic structure of a fuzzy inference system

2.2 Genetic algorithm

GA is a combinatorial optimization technique based mechanics of the natural selection process (biological evolution) of a randomly chosen population of individuals can be thought of as a search through the space of possible chromosome values or search for an optimal solution to a given problem. The basic concept is that the strong tend to adapt and survive while the weak tend to die out. The evolutionary search process is influenced by the following main components of a GA [8]: an *encoding* of solutions to the problem as a chromosome or genome; a *function* to evaluate the *fitness*; *Initialization* of the initial population; *Selection* operators; and *Reproduction* operators. During each temporal increment, the structures in the current population are rated for their effectiveness as domain solutions (called a generation). GA has been successfully used in a wide variety of problem domains (Goldberg, 1989) [8, 11, 12, and 13].

2.3 Data mining

Data mining, "discovering hidden value in your data warehouse", is frequently described as *the process of extracting valid and actionable information from large, complex databases.* In other words, data mining derives patterns and trends that exist in data. These patterns and trends can be collected together and defined as a mining model. Mining models can be applied to specific business such as forecasting sales, determining specific customers and likely products to be sold [14, 16]. The evolved system trains the data base to classify customers in two groups *normal* and *risk*. According to some variables in specific period can define customer as normal or default. This classification helps bank to remote the system for better customer relationship management.

3. Fuzzy genetic classifier model:

In the customer classification, there are two problems that can influence its accuracy. First, in credit scoring areas bank usually cannot label one customer as absolutely good who is sure to repay in time, or absolutely bad who will default certainly. Second, database is large, often complex and incomprehensible with many irrelevant variables. These redundant irrelevant variables spoil the classification, and increase many unwanted calculations and decrease the accuracy of customer classification. Moreover, it would be highly desirable to be in possession of a so-called *degree of confidence*: the system not only provides a binary classification (normal or risk), but also outputs a numeric value that represents the degree to which the system is confident about its response. Second, it would be highly beneficial for such a classification system to be human-friendly, exhibiting so-called *interpretability*. The proposed method combines two methodologies fuzzy systems and genetic algorithms, by using a simple GA with binary coding, to produce new generation of fuzzy rules based on a database of training cases (fitness) and formation of gene pool for extract good rules. The binary coding will represent all the membership functions associated to the linguistic labels belonging to each one of the linguistic term sets into a single chromosome, where are shifted along the x-axis freely. All the individuals in the population will represented by chromosomes with fixed same length. The individual (rule) will represented by the following form as Eq. (4):

$$If\ V_1\ is\ A_{j1}\ and\ V_2\ is\ A_{j2}\ and\ \ldots\ and\ V_i\ is\ A_{ji}\ then\ Y\ is\ B_K \qquad (4)$$

Where the linguistic labels A_i and B_K associated to the linguistic variables V_i and Y respectively, where $i = 1, \ldots, n, k = 1, \ldots, m, j = 0, \ldots, 3$ where n is the number of the variables. As shown in figure (“2”), fuzzy variable *u* with two possible fuzzy values labeled *Low* and *High*, so the associated membership function are two points *(p, d)*, each point is represented by a binary number with a fixed number of bits determined by u_n which is the number of possible values for the variable V_i. *P* defines the start point to measure the degree of membership versus input values, and *d* defines the length of membership function edges which separates between two labels (*Low*, *High*).

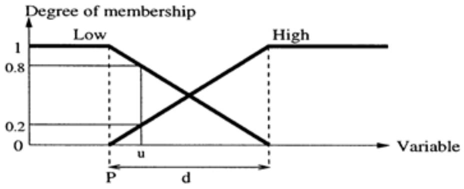


Fig. 2 example of a fuzzy variable

Each rule will be encoded in pieces of the chromosome $C_{ri}, i = 1, \ldots, n$ where n is the number of the variables in the following way:

$$C_{ri} = (p_{i1}, d_{i1}, A_{i1}, \ldots\ldots, p_{in}, d_{in}, A_{in}) \qquad (5)$$

The proposed model for classification problem consists of a fuzzy system and threshold system as shown in figure (“3”). The fuzzy system computes a continuous appraisal value of the risky customers, based on the input values. The threshold unit then outputs a normal or risk classifier according to the fuzzy system’s output.

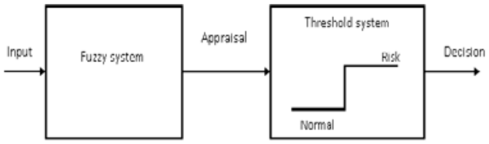


Fig.3 The proposed classifier system, note that the fuzzy subsystem displayed in Fig. 1

The proposed algorithm used to search for three parameters the relevant variables, the input membership function values, and the antecedents of rules. They are constructed as follows:

- Membership function parameters. There are *i* variables (V_i–V_n), each with two parameters *P* and *d*.
- Antecedents. The *i-th* rule has the form:

if** (*V_1 ***is A_{j1} *) **and…and** (*V_8 ***is*** A_{j8} *) **then** (output **is** normal)*

So A_{ij} represents the membership function applicable to variable V_i. A_{ij} takes the values: 1(*Low*), 2 (*High*), and 0 or 3 (*Other*).

- Relevant variables are searched for reduction by letting the algorithm chooses number of existent membership functions as valid antecedents; in such a case, the respective variables is considered irrelevant. For example, the rule

if (V_1 **is** High) **and** (V_2 **is** Other) **and** (V_3 **is** Other) **and** (V_4 **is** Low) **and** (V_5 **is** Other) **and** (V_6 **is** Other) **and** (V_7 **is** Other) **and** (V_8 **is** Low) **then** (output **is** normal),

is interpreted as:

if (V_1 is High) and (V_4 **is** Low) **and** (V_8 **is** Low) **then** (output **is** normal).

The parameters encoding are shown in figure ("4"), which form a single individual's genome. For each V_i have two points (p, d) with the same number of bits for each one and relevant variable (A_{ij}) where $i = 1, \dots, n$ and $ɉ = 0, \dots, 3$. Each chromosome has a fixed length for all individuals can be calculated according to following equation:

$$\ell_{\mathbf{c}} = (\ |p| + \ |d| + \ |A|) * n \qquad (6)$$

n is the number of variables which represented in the chromosome in bits.

P_1		d_1		A_1		……		P_4		d_4		A_4		……		n		……	
1	0	1	0	0	1	1	1	0	0	1	1	1	0	0	0	1	0	1	1

Fig. 4 encoding of the chromosome

The structure of the chromosome generated randomly with initial database that used to tune the database parameters (*p*,*d*) and rule base constituted by *m* control rules.

3.1 Cluster explanation:

The proposed method starts with a set of solutions to the problem under examination, the solutions set (represented by chromosomes in GA) is called the population which generated randomly. Every evolutionary step, known as a generation, the individuals in the current population are decoded and evaluated according to some predefined quality criterion, referred to as the fitness, or fitness function.

Each individual generates a fuzzy rule. This rule is trained by the database values which presented as input values to the fuzzy system. The membership value of each variable is then computed as $\mu Low(u)$ and$\mu\ high(u)$, as shown in figure ("2"). Therefore, the inference engine goes on to compute the truth value of each rule by applying the fuzzy logic operator (i.e. complement, intersection, union) to combine the antecedent clauses (the membership values) in a fuzzy manner (1), this results in the output truth value, namely, a continuous value which represents the rule's degree of activation [15]. The defuzzifier producing the final continuous value of the fuzzy inference system; this latter value is the appraisal value that is passed on to the threshold unit, Figure ("3") which calculated as follow:

$$Appraisal = \frac{W(a) * \mu A(u) \vee \mu B(u) * B_{min} + w(d) * B_{max}}{W(a) * \mu A(u) \vee \mu B(u) + W(d)} \qquad (7)$$

where $W(a)$ and $W(d)$ are weights of the active rule and default rule, respectively. The continuous appraisal value would be in the range of (B_{min} ,B_{max}) discrete values of output membership function. This value is then passed along to the second subsystem *threshold subsystem* which produces the final binary output (B_{min} or B_{max}). The threshold subsystem simply the outputs B_{max} if the appraisal value is above a fixed threshold value and outputs B_{min} otherwise as illustrated in figure ("3"). The threshold value can be rewritten as follows:

$$\theta = (B\_min + \ B\_max)/2 \qquad (8)$$

Then the system computes the fitness function $\delta_{a^{\kappa}}$; means the percentage of cases correctly classified by the following equation:

$$\delta_{a^{\kappa}} = \sum_{E \in cases}^{0} ((d_{\kappa}(E) - a_{\kappa}(E))^2) \qquad (9)$$

where the $k = 1, .., m, m$ is the output decision for the rule and the $d_{\kappa}(E)$ is the desired output for the case and $a_{\kappa}(E)$ is the actual output of case estimated by the system. The proposed model uses genetic algorithm with a fixed population size of μ individuals to evolve the fuzzy inference system, and fitness-proportionate selection (higher fitness more likely) and genetic operators. To form a new population (the next generation), individuals are selected according to their fitness. Thus, high-fitness ('good') individuals stand a better chance of 'reproducing', while low-fitness ones are more likely to disappear. The evolved system puts these individuals in mating pool to generate good children by using genetic operators. Crossover operation is used to obtain a new individual by combining different chromosomes to generate new better child using crossover operator *pc*, the new solution carried out by flipping bits at random; with usually small probability *pm* [12]. The evolved system made it changeable to get optimized solution and small to ensure that the good solutions are not distorted too much. In each iteration, the evolved system runs the generating method for choosing the best chromosome. The algorithm terminates when the maximum number of generations is reached.

3.2 The fuzzy genetic classifier algorithm*:*

```
Input: training set, control parameters
Output: the optimized rule set
Begin:
1) Initialize control parameters;
2) t=0 , generation counter
3) Generate initial population randomly, $C(0)$, of $\mu$ individuals;
    for each individual, $X_i$ (t) ∈ C(t) do
            Evaluate the fitness, f ($X_i$ (t));
    End
4) For each generation

        a)  Choose 2 parents at random;
        b)  Create offspring $X'_i$ (t) through application
            of crossover operator on parent genome
        c)  Mutate offspring according to mutation
            operator;
        d)  Evaluate the fitness of offspring, f($X'_i$(t))
        e)  If  f($X'_i$ (t))  >  f($X_i$ (t))  then
                add $X'_i$ (t) to c (t+1)
            Else    add $X_i$ (t) to c (t+1),


5) Select the individuals to form rule set, which satisfy the higher fitness;
 End
```

4. Results

According to the database which was collected from Barclays bank in the period from 30/12/2008 to 30/5/2009 as a six month used for training data the new model exhibits classification of the bank's customers in two groups normal and risk. The database consists of eight measured variables (V_1-V_8) as follow :(Average Revenue as V_1, Internal Transfer as V_2, Foreign Transfer as V_3 , Loan Count as V_4 , Loan Over Due as V_5 , Guarantees as V_6, Insurance as V_7, CC _ Over Due as V_8) as shown in table("1").

Table 1: customer classification

Customer id	V1	V2	…….	V8	Decision
100	1300	8	……	5	Normal
101	3000	6	……	10	Risk
……	……	…..	……	…..	……
970	450	6	……	3	Normal

The fuzzy system setup

<u>Logical parameters</u>
- Reasoning mechanism: singleton-type fuzzy system.
- Fuzzy operators: min.
- Input membership function type: orthogonal.

<u>Structural parameters</u>
- Number of input membership functions: two membership functions denoted *Low* and *High*.
- Number of output membership functions: two singletons are used, corresponding to the *normal* and *risk* classifier.
- Number of rules: The rule itself is to be found by the genetic algorithm.

<u>Connective parameters</u>
- Antecedents of rules: to be found by the algorithm.
- Consequent of rules: the algorithm finds rules for the *normal* class; the *risk* class is an else condition.
- Rule weights: active rules have a weight value 1 and the else condition has a weight of 0.25.

<u>Operational parameters</u>
- Input membership function values: to be found by the evolutionary algorithm.
- Output membership function values: we used a value of 3 for *normal* and 5 for *risk*.
- Threshold value: 4.

The evolutionary algorithm setup

- $\ell_{\mathbb{C}}$: length of the chromosome = 124 bit.
- *Selection method*: roulette wheel selection.
- *Population size*: 200.
- *Crossover operator*: 0.60.
- *Mutation operator*: 0.032.

The evolutionary performed into eight categories according to partitioning data into two distinct sets training set and test set as shown in table ("2"); The table lists the average performance over all 100 evolutionary runs, where the averaging is done over the best individual of each run. The performance value denotes the percentage of cases correctly classified. Three such performance values are shown, (1) performance over the training set; (2) performance over the test set; and (3) overall performance, considering the entire database. The choice of the training set is done randomly. The number of rules per system was fixed between one and six determined by the final structure of the genome.

Table 2: results summary of 100 evolutionary runs

Training/test cases	Performance training set	Test set	Overall
870/0	-	-	97.2
750/120	97.2	96.4	96.8
650/220	97.4	96	96.7
550/320	97.7	94.7	96.2
450/420	97.9	94.1	96
350/520	98	93.8	95.9
250/620	98.3	93.3	95.8
150/720	98.6	92.8	95.7

Figure ("6") shows the accuracy of the proposed model according to number of used cases in the training, the classification rate accuracy increases with the number of cases used after 100 evolutionary runs.

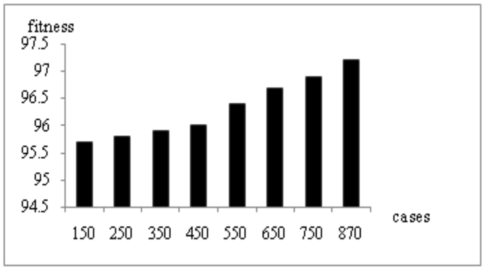


Fig. 6 classification rate

Finally, table ("3"): delineates the best one-rule system found through proposed evolutionary approach with its initial database and its rule base. It obtains 97.2% correct classification rate overall the customer cases.

Table 3: the best evolved fuzzy classification system with one rule

<table>
<tr><td colspan="9">Data base</td></tr>
<tr><td></td><td>V1</td><td>V2</td><td>V3</td><td>V4</td><td>V5</td><td>V6</td><td>V7</td><td>V8</td></tr>
<tr><td>P</td><td>4773</td><td>9</td><td>2</td><td>2</td><td>1</td><td>39</td><td>7736</td><td>1</td></tr>
<tr><td>d</td><td>7257</td><td>18</td><td>6</td><td>2</td><td>5</td><td>95</td><td>6593</td><td>7</td></tr>
<tr><td>A</td><td>3</td><td>2</td><td>1</td><td>0</td><td>1</td><td>2</td><td>0</td><td>3</td></tr>
<tr><td colspan="9">Rule base</td></tr>
<tr><td colspan="2">Rule</td><td colspan="7">If (v2 is high) and (v3 is low) and (v5 is low) and (v6 is high)then (output is normal)</td></tr>
<tr><td colspan="2">default</td><td colspan="7">Else (output is risk)</td></tr>
</table>

[18] Uses fuzzy decision trees for classification of customers' problems by automatically creating fuzzy regions around tree nodes. [19] Proposes a fuzzy 'if-then' rule based classifier to predict bankruptcy in banks. After comparison with the proposed method on the data base of Barclays bank the following results achieved as shown in the figure ("7"). The proposed method (F.G) achieved higher classification accuracy on the data base than the other previous methods.

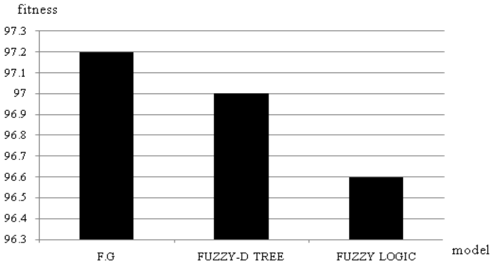


Fig. 7 comparison between classification rates

5. Conclusion

The paper explains the proposed algorithm (hybrid fuzzy- genetic algorithm) to solve the bank customers' classification problem. Proposed evolved fuzzy system presents both characteristics, first: attain high classification performance with possibility of measure to the output, second the results provide simple rules with its interpretable. Proposed model powerful in financial treatments because it could determine if the customer good or otherwise, with few determined measured variables. Fuzzy-natural computing hybrid techniques have been successfully applied to several fields of soft computing for intelligent data mining. For future work will add artificial immune systems as a natural computing technique with fuzzy logic for better classification.

Reference

[1] A. E. Elalfi, R. Haque, and M. E. Elalami, "Extracting rules from trained neural network using GA for managing E-business", Applied Soft Computing, Vol. 4, NO. 1, February 2004, pp. 65-77.

[2] F. Hoffmann, B. Baesens, and J. Martense, "Comparing a genetic fuzzy and a neurofuzzy classifier for credit scoring", international journal of intelligent systems, vol. 17, NO. 11, 2002, pp. 1067-1083.

[3] Z. Zhang, L. Zhang, and Sh. Niu, "A Parallel Classification Algorithm Based on Hybrid Genetic Algorithm", IEEE international conference on Intelligent control and automation, 2006, vol. 6, pp.3237-3240.

[4] X. Chuansheng, X. Xin, and H. Wentian, "Power Customer Credit Rating Based on FCM and the Differential Marketing Strategy Research", information science and engineering (ICISE), 2010, 2nd international conference,17 January 2011,vol. 2, pp.416-418.

[5] H. Wang, and Y. Xiang, "Study on Customer Classification Based on BP Neural Networks", IEEE international conference on Wireless Communications, Networking and Mobile computing, 2008, WiCOM'08, 4th international conference, 18 November 2008, vol. 4, pp 1.

[6] C. Chiu, "a case based customer classification approach for direct marketing", expert systems with applications, February 2002, vol.22, pp.163-168.

[7] G. Yang, and X. Yuan, "bank customer classification model based on Elman neural network optimized by PSO", IEEE international conference on Wireless Communications, Networking and Mobile computing, 2007, pp. 5672-5675.

[8] A. P. Engelbrecht, "Computational Intelligence: An Introduction", 2 edition, Wiley, England, 2007.

[9] F. Herrera, M. Lozano, and J. Verdegay, "Generation fuzzy rules from examples using genetic algorithms", fifth international conference of information processing and management of uncertainty in knowledge-based system, Paris, 1994, Vol. 5, pp 675-680.

[10] D. Fasel, "a fuzzy data warehouse approach for the customer performance measurement for a hearing instrument manufacturing company", IEEE international conference on Fuzzy Systems and Knowledge Discovery, 2009, Vol. 6, pp. 285-289.

[11] R. R. Yager, and D. P. Filev., "Essentials of Fuzzy Modeling and Control", SIGART Bulletin, Vol.6, NO. 4, 1994.

[12] O. Ahmed, M. Nordine, S. Sulaiman, and W. Fatimah, "Study of Genetic Algorithm to Fully-automate the Design and Training of Artificial Neural Network" , IJCSNS International Journal of computer Science and Network security, Vol.9 No.1, January 2009,pp. 217-226

[13] P. Makvandi, J. Jassbi, and S.Khan, "Application of Genetic Algorithm and Neural Network in Forecasting with Good Data", WSEAS International Conference on neural network, Lisbon, Portugal, 2005, Vol. 6, pp.56-61.

[14] J. Han, and M. Kamber, "Data mining: concepts and techniques", 2nd Edition, Morgan Kaufmann, 2006.

[15] C.A. Pena Reyes, and M.A. Sipper, "fuzzy-genetic approach to breast cancer diagnosis", Artificial Intelligence in Medicine, Vol. 17, No. 2, 1999, pp. 131-155.

[16] A. Berson, and Kurt Thearling, "Building Data Mining Application for CRM", USA, 1999.

[17] Z. Michalewicz, "Genetic Algorithms + Data Structures=Evolution Programs", 3rd edition. Berlin Heidelberg, Springer-Verlag, inc., 1996.

[18] K. Crockett, and Z. Bandar, "Fuzzification of Discrete Attributes From Financial Data in Fuzzy Classification Trees", IEEE International Conference on Fuzzy Systems, Korea, 2009, Vol. 18, pp 1320-1325.

[19] P.R. Kumar, and V. Ravi, "Bankruptcy Prediction in Banks by Fuzzy Rule Based Classifier", IEEE International Conference on Digital Information Management , 2006, Vol. 1,pp 222-227.

Feature Selection for Generator Excitation Neurocontroller Development Using Filter Technique

Abdul Ghani Abro [1], Junita Mohamad Saleh [2]

School of Electrical & Electronics Engineering
Engineering Campus, Universiti Sains Malaysia
14300 Nibong Tebal, Seberang Perai Selatan
Penang, Malaysia

Abstract

Essentially, motive behind using control system is to generate suitable control signal for yielding desired response of a physical process. Control of synchronous generator has always remained very critical in power system operation and control. For certain well known reasons power generators are normally operated well below their steady state stability limit. This raises demand for efficient and fast controllers. Artificial intelligence has been reported to give revolutionary outcomes in the field of control engineering. Artificial Neural Network (ANN), a branch of artificial intelligence has been used for nonlinear and adaptive control, utilizing its inherent observability. The overall performance of neurocontroller is dependent upon input features too. Selecting optimum features to train a neurocontroller optimally is very critical. Both quality and size of data are of equal importance for better performance. In this work filter technique is employed to select independent factors for ANN training.

Keywords: *neural network, mlp, feature selection, regression analysis, generator excitation*

1. Introduction

In recent years it has been recognized to impart more flexible control systems, it is necessary to incorporate other elements, such as course of thoughts, reasoning and heuristics into algorithmic techniques of conventional adaptive and optimal control theory. For proper designing of adaptive controller flexibility is main characteristic to incorporate and Artificial Neural Network (ANN) offers highly flexible structure. The use of an ANN with its learning ability avoids complex mathematical analysis in solving control problems when plant dynamics are unpredictably complex and highly non-linear [1]. This is a distinctive advantage over the traditional non-linear control methods.

ANNs are parallel distributed processing systems capable of synthesizing a complex and highly nonlinear mapping from input feature space to output space [2]. The parallel processing element distribution not only gives higher degree of tolerance but also the capability of fast information processing. Another important feature of ANN is learning and adaptation. A well trained ANN has the ability to generalize training pattern. In addition to their ability to produce high quality results for large, noisy or incomplete data sets, ANNs have been found effective in identifying patterns and other underlying data structures in multidimensional data [3].

Importance of input variables is evident as the input vector needs to capture all characteristics of complex functions. Features, variables, attributes, parameters are used alternatively for input vectors given to ANN. Feature selection is a problem of selecting the subset of features that is needed to describe the target concept in a give data set, indeed [4-6]. Keeping this in mind, one can say that it is not necessary that best individual feature correspond to best set of feature. Therefore, for best subset of features researcher better undergo all possible combinations of features available in feature set for optimum efficiency.

Feature selection is fundamental because it allows us to reduce the various effects causing information overlapping, noise induction, highly complex computation, cost of computation, memory requirement, time to compute and inter variable correlation [7]. Alternatively, too few features may carry very low content information and too many may cause irrelevant features, complex mapping and data over fitting [3]. However, it should be pointed out that a larger number of training data should always be favored as opposed to smaller number. Hence, it is issue of harmonizing irrelevant data and information.

There are two techniques employed for feature selection. One is filter-based approaches employing statistical tests for feature selection. Another is called wrappers methods exploit the knowledge of the specific structure of the learning algorithm and cannot be separated from it [8]. In the absence of valid and reliable evaluation, there currently exists no consensus on which methodology should be applied under which data condition [9]. Wrapper technique based feature selection is computationally very expensive [7] and this is big flaw behind not being used so frequently. In comparison filters based feature selection methodology is most frequently used for efficient feature evaluation [9]. In this research work forward selection based on statistical methods is used to select optimum sub set of features.

Since the discovery of Multi Layer Perceptrons' (MLP) nonlinear problem solving ability, there has been an explosive growth in application of ANN into control problems. MLP is most commonly used ANN topology and type is a type of feed-forward network. MLPs are finding more applications because of their simplicity and requirement of lesser features to approximate any function up to same degree of certainty. It contains one or more hidden layers. The number of nodes in the hidden layers defines the complexity and the power of the neural network model to describe underlying relationship and structure inherent in training data [3]. ANNs have a specific nonlinear function associated with every number of hidden layer size. However, it cannot be interpreted because of poor interpretability of ANN. Generally, one hidden layer with sigmoidal activation function is used with sufficient accuracy to approximate any nonlinear function. A key challenging aspect of the MLP-ANN is the optimization of network training protocols that include network architecture and training stopping criterion [10].

Due to the nonlinear and highly dynamic nature of power system and complexity involved in realization of optimal and nonlinear controllers, artificial intelligence particularly ANNs are finding wide variety of applications in operation and control of power system [11-17]. There are thousands of papers published in this area but literature reviewed here is as per scope of this paper focusing on excitation of synchronous generator. Work proposed by [18] has used Functional Link Net (FLN) and technique researched in [19] has proposed a method equivalent to conventional self tuning adaptive control utilizing RBF feedforward network. Research proposed in [1] , [20] and [21] is based on indirect adaptive control, utilizing three layer MLP to realize model and neurocontroller. Adaptive Critic Design (ACD) based control utilizes Hamilton–Jacobi–Bellman equation based optimal control algorithm. Duel Heuristic Programming (DHP) based ACD has been shown to perform better [22]. Work proposed by [23] use MLP based critic control, whereas MLP and RBF based critic control comparison was carried out in [24]. RBF showed better performance for low magnitude disturbances. In [25] by using RBF based adaptive critic neurocontroller, it is showed that performance of neurocontroller is better even when conventional excitation system is equipped with power system stabilizers (PSS).

In indirect adaptive based control, since link between current system state and the controller parameters are totally ignored [26] and the identified model has error of considerable percent [27], then may controller generate erroneous signal and lead to oscillatory response. Whereas in ACD, complicated control algorithm needs more computational time to calculate control signal [28] and response time is key to close loop control system performance specifically dynamic system such as power system. Additionally, reliability of ACD based control loop is also low. On the contrary, if not impossible at least it is time-consuming to train neurocontroller offline for every operating condition.

Generalization means to capture trend in data instead of fitting every training data set. Alternatively, close inputs ought to generate close outputs. For better generalization early stopping criterion plays very important role. Apart from that, early stopping of ANN training saves training time. However, if criterion to stop training is not appropriately chosen that may lead to under trained network. In this research work, ANN training was stopped on the basis of the network's performance on validation data set. This early stopping criterion is not used in the realm of power system control and operation. This is explained in the last section.

As explained, a highly challenging task to train ANN for power system control and operation is selection of input features. Aforementioned literature review reveals that variables given in Table 1 were used for ANN training to control excitation of synchronous generator. The output of the excitation system is called excitation voltage (V_f) and it is a dependent parameter. Detailed explanation of excitation system's impact on generator operation is given in next section.

Table 1 Input feature used in ANN training

ΔV_T	Terminal Voltage
ω	Rotor Speed
P	Active Power
Q	Reactive Power

ΔV_T is deviation of terminal voltage from reference voltage i.e. V_{REF}-V_T

No proper procedure has been reported for selecting input features for generator excitation neurocontroller training. More input features may require many processing elements and hence more information processing time. On the other hand, multicollinearity between input features may inhibit a neurocontroller's learning capability. In addition, un-correlated input and output space make the mapping very complex. Furthermore, generator terminal voltage can be sensed by different combinations of Table 1 parameters and even with few additional factors. This analysis is first of its own kind, to the authors best knowledge such analysis was not carried keeping statistical and engineering constraint both at a time.

Statistical methods and ANN have been used for prediction and approximation, with ANN giving higher accuracy in high dimensional problems. In fact, the most commonly used ANN topology, called multilayer perceptron is nothing more than nonlinear regression. By using statistical methods optimum parameters can be found [29] for enhancing neuro controller learning capability while generator dynamics remain unaffected. Objective of this paper is to generate optimal set of training features and to compare performance of statistical regression and ANN.
This paper is divided into three parts. The immediate section discusses the model considered for data generation. Second segment describes data analysis based on statistical methods and last part focuses on ANN output and comparison.

2. Power System Modeling

Power system is spread over very wide region from one end of country to another end and sometime from one continent to another, comprising of generation, transmission and distribution sections. The primary element of generation section is synchronous generator also termed as alternator. Synchronous generator consists of stator called armature and rotor also known as field. Field is responsible for keeping air gap magnetic flux constant leads to constant terminal voltage. The key to proper operation of synchronous generator is maintenance of synchronism between rotating armature flux and revolving field flux. The strength of synchronism largely depends upon the strength of air gap magnetic flux and alternatively dependent upon excitation system performance. Synchronism can be jolted by faults induced anywhere in a power system, but the extreme disturbance is fault introduced at terminals of a generator. Fault deteriorates the strength of magnetic flux as explained by armature reaction phenomenon and so has the effect on synchronism. The mechanical angle between rotor magnetic field and armature magnetic flux of a generator is known as the load angle or power angle, δ.

The ability of power system to regain a state of operating equilibrium after being subjected to a physical disturbance or fault is called power system stability. In addition, neither a unit at generating station nor a portion of power system should lose synchronism with respect to the generating station or the power system [30]. Power system stability enhancement has captured growing attention of researchers in recent times after occurrence of major blackouts [31].The excitation system's output is based on the difference (ΔV) between reference voltage and terminal voltage. The fault causes a decrease in air gap flux density, depending upon the direct and quadrature axis sub-transient and transient time constant. Moreover duration of fault, and decrease in terminal voltage have great influence on air gap flux density reduction. This leads to increase in ΔV, so the output of generator excitation will shoot up to compensate error. Stability may be enhanced by rapidly increasing excitation current [32]. ANN requires quite considerable time to tune weights but it is fast and accurate once tuned properly. In this research work besides variables given in Table 1, the effect of one more new variable, deviation of quadratic voltage from reference voltage i.e. $\Delta V_q = V_{ref} - V_q$ was analyzed on excitation voltage (V_f). Terminal voltage is the vector sum of direct and quadratic voltage components. Quadratic voltage was preferred over direct voltage because of higher correlation constant. The reference value for quadratic voltage is achieved by putting one instead of terminal voltage in equation combining direct, quadratic and terminal voltage.

Power system stability enhancement is referred to reducing risk of losing stability by inserting additional signals into the system to smooth out the system dynamics. During steady state excitation system should be driven by only voltage difference. Contrastingly, during transient state rotor swings ΔV undergoes oscillations caused by change in rotor angle. It is compulsory to add additional information to neurocontroller for damping out oscillations. Rotor speed, active power or both are usually used variables for generating stabilizing signals [11, 33, 34]. In this research work one more parameter load angle (δ) is also included for analyzing learning performance based on its correlation with excitation voltage and active power. Selection of load angle will not affect negatively the generator dynamics because active power and load angle are proportional as evident from equation

$$P = \frac{E_f * V_T}{X_S} * \sin\delta \tag{1}$$

where P is active power, E_f is internal generated voltage, V_T is terminal voltage, X_S is synchronous reactance of generator and δ is load angle. Single machine infinite bus system (SMIB) power system model is considered for generating data, as shown in Figure 1. This model simulates a generator connected with the rest of power system.

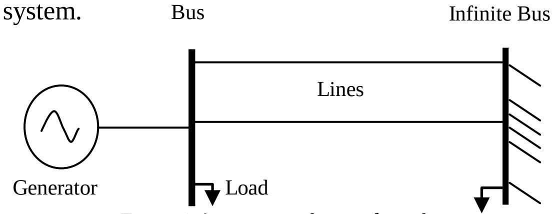


Figure 1 A single machine-infinite bus system

Simulation of the model was carried out on Matlab/Simulink Toolbox with generator rating 13.8KV, 150MVA, 50Hz at load (0.09+j0.056) Ω. The generator parameters and excitation system parameters are given in Table 2 and 3 respectively. A three phase to ground fault was simulated to analyze system transient stability. Figure 2 and 3 show terminal voltage and load angle behavior after simulation of 120ms fault at generator terminals. Both figures depict stable behavior of the generator. This implies data were collected from a stable system.

Table 2 SYNCHRONOUS GENERATOR PARAMETERS

X_d	=	1.83	X_q	=	1.7	R_{Stator}	=	0.003
X'_d	=	.24	X'_q	=	0.43	Inertia	=	3.6
X''_d	=	.20	X''_q	=	0.26	Hz	=	50
T'_d	=	0.3s	T''_d	=	0.04s	T''_q	=	0.031s

Table 3 EXCITATION PARAMETERS

Ka	=	2.50	Ta	=	0.001s
Ke	=	1.5	Te	=	0.3s
Kf	=	1	Tf	=	0.003s

The field of statistics deals with the collection, presentation, analysis and use of data to make decisions. Statistical methods are used to assist for describing and appreciating variability. Variability means the successive observations of a system or phenomenon do not always produce exactly the same result. Hence statistical thinking gives us a useful way to incorporate variability into decision making process.

Statistical analysis was carried out using Minitab software. In statistical modeling data generation play an important role in model acceptance. In this work aforesaid model simulation include ±10% change in Vref, self clearing and not self clearing three phase to ground fault at generator terminals and transmission line tripping and addition as the types of disturbances. Data were sampled at 200Hz sampling frequency. Then fifty random samples were taken for further analysis. Care was taken that the sample should be a true representation of whole population space

2. Data Analysis

Since power system requires high degree of reliability, hence fewer inputs are preferred to use. The efficiency of control signal is increased by using time delayed values as power system is dynamic system. This is the main reason why this analysis did not use stepwise regression analysis and only relied on linear regression. The statistical modeling process involved three steps [35]: (i) correlation analysis, (ii) regression analysis, and (iii) model assessment [36]. These steps are discussed below.

(a) Correlation is the process for determining the strength of relation between dependent and independent variables. Table 4 shows Pearson correlation between various independent parameters and the dependent factor, excitation voltage (V_f). The table also shows significance, called probability value (P-value). Pearson correlation was chosen because the data is scaled type, i.e. value varies from -∞ to +∞.

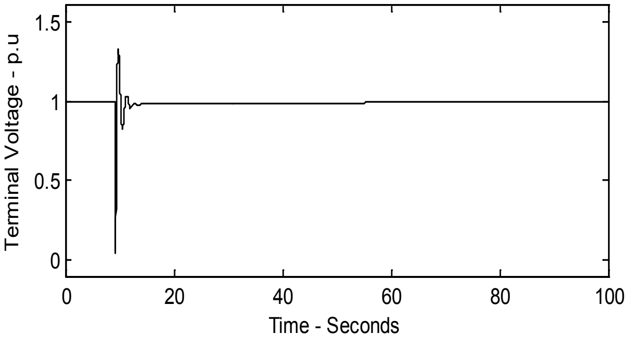


Figure 2 Terminal voltage after 120ms fault at generator terminals

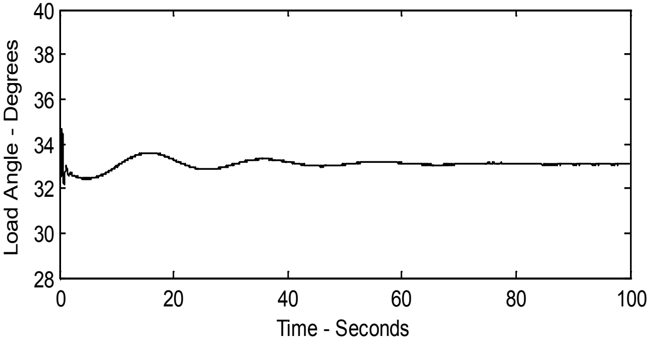


Figure 3 Load angle (δ) transition after fault

Table 4. Statistical Correlation Test Output

Independent Variables	Correlation coefficient	Significance P-Value
ΔV_T	0.587	0.000
ω	-0.06	0.486
P	0.648	0.000
Q	0.635	0.000
ΔV_q	0.759	0.000

Correlation is between Excitation Voltage (V_f) and parameters given.

For 95% confidence level, the significance of less than 0.05 considered statistically meaningful. The table suggests strongest correlation between quadratic voltage (**V_q**) and excitation voltage (**V_f**), followed by active (**P**) and reactive power (**Q**). The results suggest no relationship between rotor speed and excitation voltage and this parameter is currently being used as auxiliary stabilizing signal. Keeping in view correlation results rotor speed is eliminated from further analysis.

(b) *Regression Analysis* gives the prediction of dependent variable based on independent factors of an empirical model. The regression equation is given as below [37],

$$\Upsilon = \beta o + \sum_{i=1}^{n} \beta i \mathrm{X} i + \varepsilon \qquad (2)$$

where β are constants and X is the independent variable and Y is the dependent variable. The random error term ε is assumed to have zero mean, constant variance σ^2 and normally distributed [38].

The accuracy of regression model is determined by coefficient of multiple determination i.e. R^2. Higher R^2 value indicates good prediction accuracy of model. Nevertheless, using only R^2 is not always a good indicator of model adequacy. Seeing that R^2 increases with addition of another regressor variable irrespective of whether additional variable is statistical significant or not. The adjusted coefficient of multiple determination i.e. R^2(Adj) is a better reflection of the model adequacy along with R^2. R^2(Adj) will increase only when additional factor is statistically significant. Lower standard deviation (S) is also conceived an indicator for better performance of the model. The low S means data set tends to be very close to mean and assumed mean in regression definition is zero.

Table 5 gives the regression analysis results of the models without having stabilizing signal. Model 1contains ΔVq and model 2 comprises of ΔV_T. Coefficient of multiple determination (R^2) of model 1 is higher than model 2. Therefore, it can be concluded that ΔV_q has higher prediction accuracy than ΔV_T. Hence it is expected, neuro controller trained on model 1 may give less error than model 2. Value of R^2(Adj) is higher for model 1 than model 2. Since in this analysis each model contains only one element R^2(Adj) does not serve its purpose here. However, this value is given here to compare in next step when stabilization signals are added and compared in different combinations.

Nonetheless, only high gain excitation system can produce low frequency oscillations which ultimately lead to unstable system. Therefore, auxiliary stabilizing signals addition to excitation system is essential for stability enhancement [39]. The following explanation considers the models comprising of additional signals. Table 6 describe the regression output of models containing active power (P) and reactive power (Q) as stabilizing input feature in combination to voltage deviation signals.

After contemplating Table 6 it can be concluded that not only the additional signals stabilize the system but also increase prediction accuracy. With additional signal addition R^2(Adj) is also increased, which is another evidence to believe higher prediction accuracy of Table 6 models. However, authors are reluctant to select P and Q as stabilizing parameter owing to higher VIF factor. Variance Inflation Factor (VIF) predicts correlation among predictors. From statistical analysis perspective VIF value up to 10 is considered normal. Nevertheless, lesser VIF value means better performance, since ANNs output is highly sensitive to VIF value. In addition, active power and reactive power being electrical signals have lower response time. These quantities are very sensitive to noise in comparison to mechanical signals conceived here onward.

Table 5 Regression output of Models without Stabilizing Signals

	Model 1	**Model 2**
Constant	1.879	-0.214
ΔV_q(coefficient)	24.944	-
Significance	0.000	-
ΔV_T(coefficient)	-	-35.617
Significance	-	0.000
S	0.95	1.233
R^2	73.3	59.8
R^2(Adj)	70.2	56.1

Table 6 Performance index for Regression of Models with Stabilizing Signals

	Model 3	**Model 4**	**Model 5**	**Model 6**
Constant	2.718	7.140	4.490	3.953
ΔV_q(Coefficient)	22.441	-	17.982	-
Significance	0.000	-	0.000	-
VIF	5.6	-	6.9	-
ΔV_T(Coefficient)	-	9.653	-	19.620
Significance	-	0.251	-	0.021
VIF	-	9.8	-	8.3
P (Coefficient)	-	8.151	3.525	-
Significance	-	0.001	0.071	-
VIF	-	6.4	7.8	-
Q (Coefficient)	22.441	-	-	17.879
Significance	0.493	-	-	0.03
VIF	5.7	-	-	4.9
S	0.985	1.108	0.955	1.1839
R^2	75.5	69.0	77.2	64.7
R^2(Adj)	73.9	67.0	75.5	62.3

Table 7 also shows regression analysis output but with different stabilizing signal in combination to voltage. Higher prediction accuracy and low VIF value prophecy better performance of Table 7 models than Table 6 models. Models 7 and 8 consisting of load angle (δ) in combinations of terminal voltage (Vt) and quadratic voltage (Vq). R^2 of model containing quadratic voltage is higher than model comprising terminal voltage and value of R^2(Adj) is also higher. The S of model 8 is lower than model 7 too, shown in Table 7. However, VIF value of model 7 and model 8 is almost equal. It can be deduced from Table 7 figures, set of input parameters containing quadratic voltage has higher prediction accuracy than set consisting of terminal voltage. Hence it is anticipated, neurocontroller trained on model 8 may give less error than model 7.

TABLE 7 Performance index for Regression of Models with Stabilizing Signals

	Model 7	Model 8
Constant	-1.825	-9.230
ΔV_q(Coefficient)	-	29.697
Significance	-	0.000
VIF	-	2.4
ΔV_T(Coefficient)	36.522	-
Significance	0.000	-
VIF	2.9	-
δ(Coefficient)	0.0419	0.2995
Significance	0.744	0.005
VIF	1.6	1.5
S	1.2451	0.9073
R^2	60.9%	79.2%
R^2(Adj)	58.3%	77.9%

(c) *Model Assessment* step is carried out after regression model is developed. Acceptability and reliability of model are carried out in this step. Fitting a regression model requires few assumptions, meeting them tells credibility of the model. It is assumed while fitting data that the residuals are randomly distributed and lie within ±2. The residuals of a regression model are given by

$$e = y_{des} - y_{est} \quad (3)$$

where e is the error, y_{des} is the desired output and y_{est} is the estimated output. Analysis of residuals is helpful in checking assumption that the errors are approximately normally distributed with constant variance. For 95% confidence interval more than or equal to 95% residuals of model ought to lie within ±2 range [36].

Performance analysis of models 3 to 6, using steps suggested in model assessment depicts pretty poor picture too. Therefore, their assessment is not shown here. Only assessment comparison of Table 7 models is described here.

Model adequacy is analyzed by beholding the behavior of model residuals. Figure 4, 5, and 6 are related with residuals of model 8 whereas Figure 7, 8 and 9 are associated with residuals of model 7. Residual plots comparison, Figure 4 and 7, of both models exhibit that the residual distribution of both model is satisfactory. The points at (0.72,-4.5) and (1.29,-0.2) are not outliers, but these represent one of many different disturbances simulated during data generation phase. More than 95% residuals of model 8 lie in the range of ±2, which indicates that the assumptions of randomly distributed residual is satisfied, as shown by Figure 5 and 6. Whereas less than 95% of residuals of model 7 lie within range of ± 2, as indicated by Figure 8 and 9.

The distribution of residuals along regression fit is shown in Figure 5 and 8 for model 8 and model 7, respectively. The comparison of both plots expose that distribution of model 8 residuals is approaching normality more than model 7 residuals.

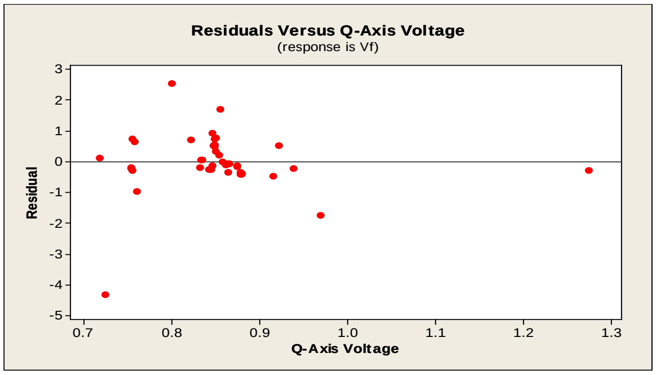


Figure 4 Residual plots of Model 8

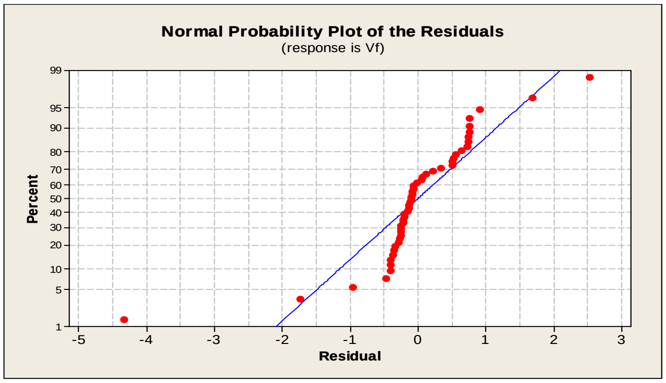


Figure 5 Normal distribution of residual of Model 8

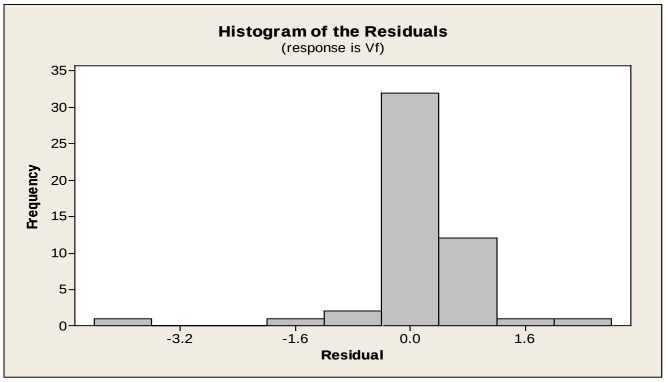


Figure 6 Histogram of Model 8 residuals

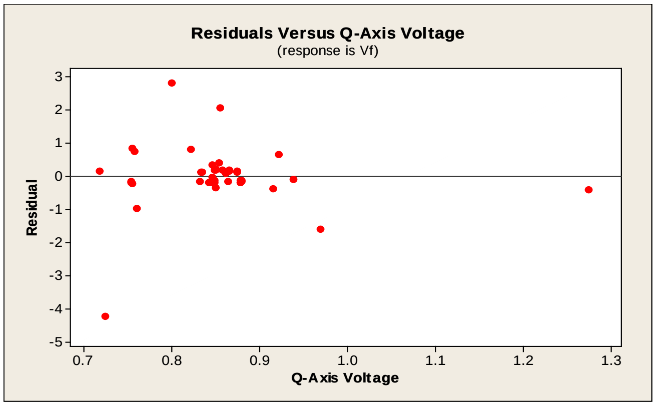


Figure 7 Residual plots of Model 7

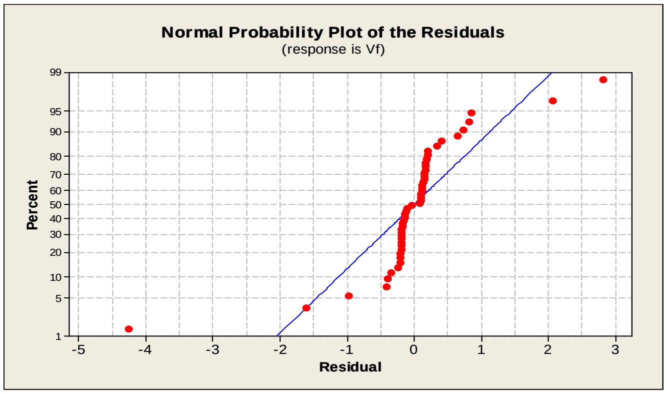


Figure 8 Normal distribution of residual of Model 7

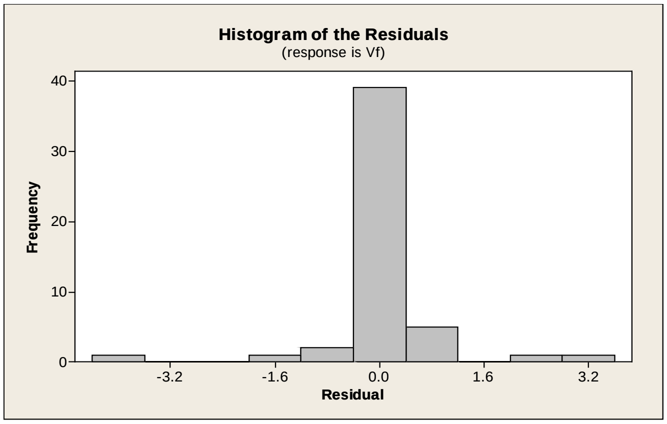


Figure 9 Histogram of Model 7 residual

4. Artificial Neural Network

In this section ANN output is discussed. In this research work MLP was chosen because of its simplicity and it is most usually used neural network [35]. A perceptron network with its adjustable hidden layer values is nonlinearly parameterized. MLP was trained using Levenberg-Marquardt Error Back Propagation. A highly challenging characteristic for a trained ANN is how well it performs when presented with new data i.e. generalization Advantages of generalization lie in adaptability, fault tolerance and model-free estimation by constructing input output mapping. As discussed in introduction, to avoid the over fitting problem, early training-terminating method called validation was employed. In this work, generated data were divided into three parts; training, validation and testing. The best MLP was selected based on one with the smallest test error. MLPs were trained on randomized data for enhancing learning capability. Basic work stages are shown in the flow chart Figure 10. The network growing technique was used to obtain an optimal MLP size. Network growing basically adds one hidden node at a time into an ANN.

MLP was trained from one to thirty hidden layer neurons, results support the selected range. Each network with each number of hidden layer neuron was trained thirty times with random initial free parameters. MLP performance was analyzed based on mean square error (MSE) and mean absolute error (MAE). The MLP with the lowest errors out of thirty run was selected for further comparison. Value of error varies in a particular fashion. Initially error value was high but decreasing with increase of hidden layer size, after touching low it starts either increasing or floating. Out of these one to thirty hidden nodes, size of neurocontroller was chosen based on minimum MAE and MSE and it is given in Table 8.

Table 8: ANN and Statistical Regression (SR) output comparison

		Model 7	Model 8
Features		ΔV_T δ	ΔV_q δ
ANN	MAE	0.277	0.245
	MSE	0.430	0.395
	HLN	23	12
	Time	0.024	0.016
SR	MAE	1.0937	0.8428
	MSE	15.4445	8.0913

SR=Statistical Regression; HLN= Hidden Layer Neurons

The ANN output is almost in proportion to statistical regression output. However the difference between errors of different sets is not in proportion to difference using statistical methods. This manifests ANN's ability to efficiently map highly complex functions. Table 8 gives the comparison of ANN output and statistical regression results. It also shows the size of hidden layer at minimum error value.

Model 8 has lower ANN error at lower hidden layer size of twelve neurons than model 7. Table 8 also depicts performance comparison of ANN, trained on both models, based on time. Model 8 has higher prediction accuracy on regression analysis and also has lesser error at lesser processing time. The lower processing time is because of smaller hidden layer size.

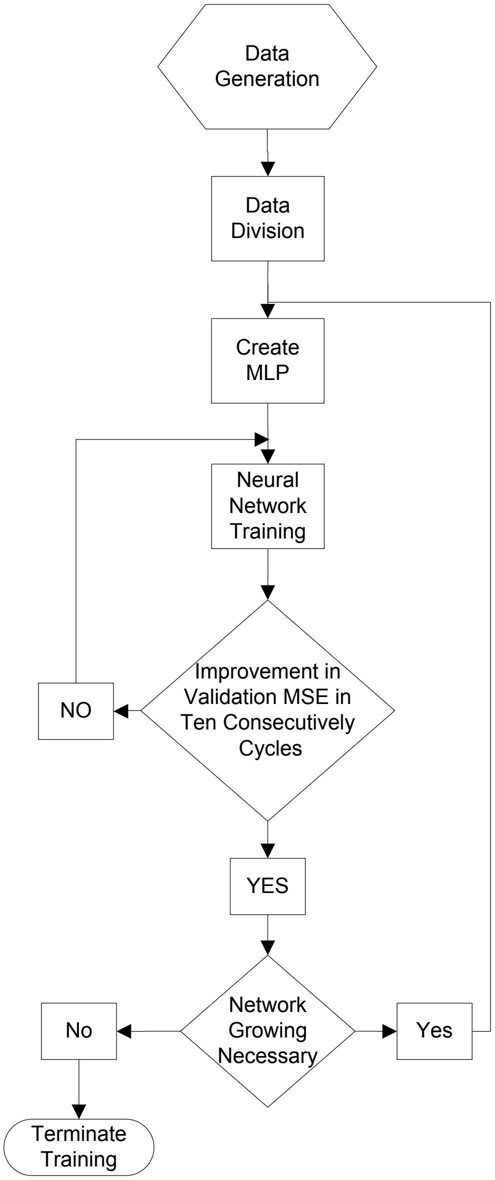


Figure 10 Flow chart showing MLP training stages

5. Conclusion

With the help of statistical analysis it is, revealed that strong correlation between input and output space enhance learning capability of ANN not only in terms of error value but also requires lesser hidden layer size. Combination of quadratic voltage and load angle (δ) is a better set of input features for synchronous generator excitation system neurocontroller training. Comparison imparts that the performance of ANN is superior to statistical regression.

Acknowledgments

The authors greatly acknowledge the support of the Universiti Sains Malaysia fellowship scheme for financial support.

Authors also want to acknowledge Associate Professor Dr Zalina Abdul Aziz for her valuable discussion and guidance from time to time.

References

1. Venayagamoorthy, G.K. and R.G. Harley, *A continually online trained neurocontroller for excitation and turbine control of a turbogenerator.* Energy Conversion, IEEE Transactions on, 2001. 16(3): p. 261-269.
2. Basheer, I.A. and M. Hajmeer, *Artificial neural networks: fundamentals, computing, design, and application.* Journal of Microbiological Methods, 2000. 43(1): p. 3-31.
3. Kavzoglu, T., *Increasing the accuracy of neural network classification using refined training data.* Environmental Modelling & Software, 2009. 24(7): p. 850-858.
4. Tirelli, T. and D. Pessani, *Importance of feature selection in decision-tree and artificial-neural-network ecological applications. Alburnus alburnus alborella: A practical example.* Ecological Informatics. (2010), doi:10.1016/j.ecoinf.2010.11.001.
5. Amit Saxena, D.P., Abhishek Dubey, *An Evolutionary Feature Selection Technique Using Polynomial Neural Network.* International Journal of Computer Science Issues, July 2011. 8(4): p. 494-502.
6. Hema Banati, M.B., *Fire Fly Based Feature Selection Approach.* International Journal of Computer Science Issues, July 2011. 8(4): p. 473-480.
7. May, R.J., et al., *Non-linear variable selection for artificial neural networks using partial mutual information.* Environmental Modelling & Software, 2008. 23(10-11): p. 1312-1326.
8. Jian-Bo, Y., et al., *Feature Selection for MLP Neural Network: The Use of Random Permutation of Probabilistic Outputs.* Neural Networks, IEEE Transactions on, 2009. 20(12): p. 1911-1922.
9. Crone, S.F. and N. Kourentzes, *Feature selection for time series prediction - A combined filter and wrapper approach for neural networks.* Neurocomputing, 2010. 73(10-12): p. 1923-1936.
10. Yang, S., G.N. Taff, and S.J. Walsh, *Comparison of Early Stopping Criteria for Neural-Network-Based Subpixel Classification.* Geoscience and Remote Sensing Letters, IEEE, 2011. 8(1): p. 113-117.
11. Nguyen, T.T. and R. Gianto, *Neural networks for adaptive control coordination of PSSs and FACTS devices in multimachine power system.* Generation, Transmission & Distribution, IET, 2008. 2(3): p. 355-372.
12. Park, J.-W., R.G. Harley, and G.K. Venayagamoorthy, *Decentralized optimal neuro-controllers for generation and transmission devices in an electric power network.*

Engineering Applications of Artificial Intelligence, 2005. 18(1): p. 37-46.

13. Amjady, N. and M.H. Velayati, *Dynamic voltage stability prediction of power systems by a new feature selection technique and probabilistic neural network.* European Transactions on Electrical Power, 2011. 21(1): p. 312-328.

14. Karami, A., *Estimation of the critical clearing time using MLP and RBF neural networks.* European Transactions on Electrical Power, 2010. 20(2): p. 206-217.

15. Farrag, M.E.A. and G. Putrus, *An on-line training radial basis function neural network for optimum operation of the UPFC.* European Transactions on Electrical Power, 2011. 21(1): p. 27-39.

16. Mazón, A.J., et al., *Strategies for fault classification in transmission lines, using learning vector quantization neural networks.* European Transactions on Electrical Power, 2006. 16(4): p. 365-378.

17. Ketabi, A., I. Sadeghkhani, and R. Feuillet, *Using artificial neural network to analyze harmonic overvoltages during power system restoration.* European Transactions on Electrical Power, 2010: p. n/a-n/a.

18. Djukanovic, M., et al., *Neural-net based coordinated stabilizing control for the exciter and governor loops of low head hydropower plants.* Energy Conversion, IEEE Transactions on, 1995. 10(4): p. 760-767.

19. Swidenbank, E., et al., *Neural network based control for synchronous generators.* Energy Conversion, IEEE Transactions on, 1999. 14(4): p. 1673-1678.

20. Venayagamoorthy, G.K. and R.G. Harley, *Two separate continually online-trained neurocontrollers for excitation and turbine control of a turbogenerator.* Industry Applications, IEEE Transactions on, 2002. 38(3): p. 887-893.

21. Salem, M.M., et al., *Simple neuro-controller with a modified error function for a synchronous generator.* International Journal of Electrical Power & Energy Systems, 2003. 25(9): p. 759-771.

22. Venayagamoorthy, G.K., R.G. Harley, and D.C. Wunsch, *Comparison of heuristic dynamic programming and dual heuristic programming adaptive critics for neurocontrol of a turbogenerator.* Neural Networks, IEEE Transactions on, 2002. 13(3): p. 764-773.

23. Venayagamoorthy, G.K., R.G. Harley, and D.C. Wunsch, *Implementation of adaptive critic-based neurocontrollers for turbogenerators in a multimachine power system.* Neural Networks, IEEE Transactions on, 2003. 14(5): p. 1047-1064.

24. Jung-Wook, P., R.G. Harley, and G.K. Venayagamoorthy, *Adaptive-critic-based optimal neurocontrol for synchronous generators in a power system using MLP/RBF neural networks.* Industry Applications, IEEE Transactions on, 2003. 39(5): p. 1529-1540.

25. Park, J.-W., et al., *Dual heuristic programming based nonlinear optimal control for a synchronous generator.* Engineering Applications of Artificial Intelligence, 2008. 21(1): p. 97-105.

26. Peng, Z. and O.P. Malik, *Design of an Adaptive PSS Based on Recurrent Adaptive Control Theory.* Energy Conversion, IEEE Transactions on, 2009. 24(4): p. 884-892.

27. Prokhorov, D.V. and D.C. Wunsch, II, *Adaptive critic designs.* Neural Networks, IEEE Transactions on, 1997. 8(5): p. 997-1007.

28. Chaturvedi, D.K., O.P. Malik, and P.K. Kalra, *Experimental studies with a generalized neuron-based power system stabilizer.* Power Systems, IEEE Transactions on, 2004. 19(3): p. 1445-1453.

29. García-Escudero, L.A., et al., *Robust clusterwise linear regression through trimming.* Computational Statistics & Data Analysis, 2010. 54(12): p. 3057-3069.

30. Kundur, P., et al., *Definition and classification of power system stability IEEE/CIGRE joint task force on stability terms and definitions.* Power Systems, IEEE Transactions on, 2004. 19(3): p. 1387-1401.

31. Esmaili, M., N. Amjady, and H.A. Shayanfar, *Stochastic multi-objective congestion management in power markets improving voltage and transient stabilities.* European Transactions on Electrical Power, 2011. 21(1): p. 99-115.

32. Wang, Y., et al., *Transient stability enhancement and voltage regulation of power systems.* Power Systems, IEEE Transactions on, 1993. 8(2): p. 620-627.

33. Abdelazim, T. and O.P. Malik. *Fuzzy logic based identifier and pole-shifting controller for PSS application*. in *Power Engineering Society General Meeting, 2003, IEEE*. 2003.

34. Chaturvedi, D.K. and O.P. Malik, *Generalized neuron-based adaptive PSS for multimachine environment.* Power Systems, IEEE Transactions on, 2005. 20(1): p. 358-366.

35. Mostafa, M.M. and R. Nataraajan, *A neuro-computational intelligence analysis of the ecological footprint of nations.* Computational Statistics & Data Analysis, 2009. 53(9): p. 3516-3531.

36. Wu, Y., Q. Zhou, and C.W. Chan, *A comparison of two data analysis techniques and their applications for modeling the carbon dioxide capture process.* Engineering Applications of Artificial Intelligence, 2010. 23(8): p. 1265-1276.

37. Young, D.S. and D.R. Hunter, *Mixtures of regressions with predictor-dependent mixing proportions.* Computational Statistics & Data Analysis, 2010. 54(10): p. 2253-2266.

38. Chaloulakou, A., M. Saisana, and N. Spyrellis, *Comparative assessment of neural networks and regression models for forecasting summertime ozone in Athens.* The Science of The Total Environment, 2003. 313(1-3): p. 1-13.

39. Chompoobutrgool, Y., L. Vanfretti, and M. Ghandhari, *Survey on power system stabilizers control and their prospective applications for power system damping using Synchrophasor-based wide-area systems.* European Transactions on Electrical Power, 2011: doi: 10.1002/etep.545

Abdul Ghani Abro received his first and second degree from MUET University – Pakistan and NED University – Pakistan. He is Assistant Professor at NED University – Pakistan and currently he is pursuing PhD studies at School of Electrical and Electronic Engineering Universiti Sains Malaysia. His research interests include application of intelligent systems for power system operation and control.

Junita Mohamad-Saleh received her B.Sc (in Computer Engineering) degree from the Case Western Reserve University, USA in 1994, the M.Sc. degree from the University of Sheffield, UK in 1996 and the Ph.D. degree from the University of Leeds, UK in 2002. She is currently an Associate Professor in the School of Electrical & Electronic Engineering, Universiti Sains Malaysia. Her research interests include computational intelligence, tomographic imaging and soft computing

An Enhanced Indexing And Ranking Technique On The Semantic Web

Ahmed Tolba[1], Nabila Eladawi[2] and Mohammed Elmogy[3]

**[1] Faculty of Computer Studies, Arab Open University
Kuwait, 3322, Kuwait**

**[2] Faculty of Computers and Information, Mansoura University
Mansoura, 35516, Egypt**

**[3] Faculty of Computers and Information, Mansoura University
Mansoura, 35516, Egypt**

Abstract

With the fast growth of the Internet, more and more information is available on the Web. The Semantic Web has many features which cannot be handled by using the traditional search engines. It extracts metadata for each discovered Web documents in RDF or OWL formats, and computes relations between documents. We proposed a hybrid indexing and ranking technique for the Semantic Web which finds relevant documents and computes the similarity among a set of documents. First, it returns with the most related document from the repository of Semantic Web Documents (SWDs) by using a modified version of the ObjectRank technique. Then, it creates a sub-graph for the most related SWDs. Finally, It returns the hubs and authorities of these document by using the HITS algorithm. Our technique increases the quality of the results and decreases the execution time of processing the user's query.

Keywords: *Indexing, Ranking Semantic Web Documents, Search Engines, Semantic Web.*

1. Introduction

The classical Information Retrieval (IR) models have been processed by using state-of-the-art models such as LSI and machine learning based models (i.e. artificial neural network, symbolic learning, and genetic algorithm) [1]. However, it has been shown that these models based on formal mathematical theories and they do not necessarily surpass the classical models [2]. In the classical IR models, matching between queries and documents is formally defined, but it is semantically imprecise. Most of these models make a plausible assumption that words in documents are independent.

On the other hand, human users are able to interpret the significance of semantic features to understand the information being presented, but this may not be so easy for an automated process or software agent. The Semantic Web aims to overcome this problem by making Web content more accessible to automated processes. The ultimate goal of the Semantic Web is to transform the existing Web into a set of connected applications and forming a consistent logical Web of data [3,4]. This can be achieved by adding semantic annotations that describe the meaning of the Web content.

Therefore, the Semantic Web will contain resources corresponding not only to media objects (such as Web pages, images, audio clips, etc.) as the current Web does, but also to objects such as people, places, organizations, and events [5]. Consequently, the Semantic Web will contain not just a single kind of relation (the hyperlink) between resources, but many different kinds of relations between the different types of resources.

This paper is divided into five sections. In Section 2, an overview of the related work, which discusses some recent ranking systems on the Semantic Web, will be introduced. Section 3 represents the architecture of our proposed system. The proposed indexing and ranking technique is described in detail. Section 4 presents the implementation and some results of our system. Finally, we conclude our work in Section 5.

2. Related Work

There are many researchers who are working on Semantic Web and how to rank the pages according to their contents. For example, TAP [6,7] was created to be an infrastructure for applications on the Semantic Web. It provides a set of simple mechanisms for sites to publish data onto the Semantic Web and for applications to consume this data. TAP improves information search and retrieval results in two ways: on the one hand, it provides a simple mechanism to help the Semantic search module to understand the denotation of the query; on the other hand, it augments the search results by considering search context and exploring closely related objects based on this context.

Kiryakov et al. [8] introduced a holistic architecture of Semantic annotation, indexing, and retrieval for documents. Their system, which is called KIM, aimed to achieve fully automatic annotation and to improve search and retrieval by integrating information extraction (IE) (i.e. using GATE [9]), information retrieval and Semantic Web technologies.

In [10-12], the authors viewed the documents representation on the Semantic Web as a combination of text, which is suitable for current Web search engines' indexing and Semantic markup. This can be used to perform inference over a knowledge-base and proposes an integrated approach to combine the inference capability and traditional information retrieval techniques. They implemented a prototype system, called OWLIR, for retrieving university event announcements.

Squiggle [13] is another framework for building domain-specific Semantic search applications. It provides capabilities for annotating, indexing, and retrieving multimedia items based upon the SKOS3 ontology.

Swoogle [14,15] is also a Semantic search engine for retrieving Semantic Web document. Its primary use is found in searching the Web and locating relevant ontologies in order to help users access, explore, and query Semantic Web documents.

Stojanovic et al. [16] have developed a domain independent approach for developing Semantic portals, viz. SEAL (SEmantic portAL), that exploits Semantics for providing and accessing information at a portal as well as constructing and maintaining the portals. They propose that the problem of Semantic ranking may be reduced to the comparison of two knowledge-bases. Query results are reinterpreted as "query knowledge bases" and their similarity to the original knowledge-base without axioms yields the basis for Semantic ranking. Thereby, they reduce their notion of similarity between two knowledge bases to the similarity of concept pairs.

Yousefipour et al. proposed an ontology-based approach for ranking Semantic Web services. A generic and domain-specific ontology is used to infer the Semantic similarity between the parameters of the request and the advertisement, which will be applied in the process of SWSs ranking. They studied how Semantic Web service ranking can be used in the context of Semantic Web service discovery.

Therefore, there are many researches on Semantic Web and how to rank the pages according to their contents. As mentioned previously, these ranking techniques do not depend only on the keywords but also on the contents of the Web documents. Consequently, the Semantic Web ranking techniques need to be developed to present an efficient way to classify SWDs and to retrieve a precise result for the user's query.

We developed an indexing and ranking technique which can be used in a Semantic Web search engine to facilitate the development of the Semantic Web and finding a proper ontology for the submitted search query. The entire documents are processed to extracted ontologies and find the relationships between these documents and the others. Therefore, our system is not based only on the extracted metadata from the document but also on extracted ontologies and the relations between documents. In other words, our main goals are to find a good measure for indexing the processed SWDs with extracting the proper ontologies, create a meaningful rank measure which reflects the importance of the processed document, and answer the user's queries efficiently.

3. System Architecture

The main architecture of our system is as shown in Fig. 1. Our system contains five components: (1) The JENA Web crawler which is used to crawl the Web and returns with SWDs to process them, (2) SWDs and metadata repositories which contain the retrieved Semantic Web Documents and their extracted metadata, (3) The Semantic ranking component which is used to index and rank the processed SWDs, (4) The pre-processing stage which stores SWDs in the repository and processes the SWDs to generate objective metadata about SWDs at both the syntax and the semantic levels, and (5) The user interface which accepts the query from the user and displays the result of the search. In the following subsections, we will discuss the components of our system in more detail.

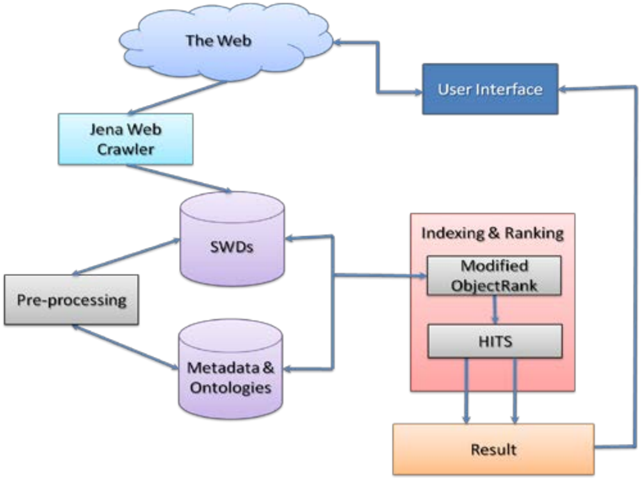


Fig. 1 The architecture of the proposed system.

3.1 Jena

Jena [18] is a Java framework for building Semantic Web applications. It provides a programmatic environment for RDF, RDFS and OWL, and SPARQL. It includes a rule-based inference engine. It is open source and grown out of work with the HP Labs Semantic Web Programme. Jena is developed to employ a number of heuristics for finding SWDs. It searches for documents of .rdf, .owl, .daml, and .n3 file extensions.

Jena analyzes the content of a SWD and discovers new SWDs. First, it verifies if a document is a SWD or not, and it also revisits discovered URLs to check updates. Secondly, several heuristics are used to discover new SWDs through semantic relations: (1) The semantics of URIref shows that the namespace of a URIref is highly likely to be the URL of an SWD; (2) The semantics of OWL shows that owl:imports links to an external ontology, which is a SWD; (3) The semantics of FOAF ontology, shows that rdfs:seeAlso property of an instance of foaf:Person often links to another FOAF document, which often is a SWD.

3.2 SWDs Repository

SWDs repository contains the Semantic Web documents which are retrieved by the Web crawler. It keeps up-to-date SWDs to use them in to answer user's query. SWDs are based on RDF which can be in RDFS, DAML+OIL, or OWL formats. They contain the following items:

- General term statements which define the classes and the properties.
- The terms' definition extensions.
- Individuals Creation.
- Make assertions about terms and individuals which are already defined or created.

Therefore, SWD can be defined as an atomic information exchange object in the Semantic Web which can be found online and accessible to Web users and software agents.

3.3 SWDs Pre-processing

The stored SWDs in the repository are processed to generate objective metadata about SWDs at both the syntax and the semantic levels. The SWDs are classified in the repository into three types: the Semantic Web ontologies (SWOs) which is called T-Boxes, the Semantic Web databases (SWDBs) which is called A-Boxes, and hybrid which defines a set of terms to be used by others as well as a useful database of information about a set of individuals.

SWD metadata is collected to make SWD processing and search more efficient and effective. SWD metadata classification can considered as a modified version of the one which is used in Swoogle. We added some additional items and changed others. Fig. 2 shows the types of SWDs metadata which are processed and stored in metadata and ontologies repository.

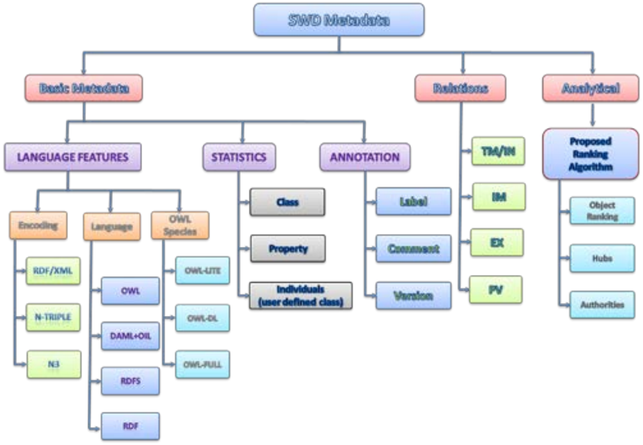


Fig. 2 Types of SWDs metadata stored in metadata &ontologies repository.

SWD metadata is derived from the content of SWD as well as the relations among SWDs. They can be classified into three categories of metadata:

- **Basic Metadata**: It considers the syntactic and semantic features of a SWD. It contains the following types:
  - Language feature: It refers to the properties describing the syntactic or semantic features of a SWD. It captures the following features:
    - Encoding: It shows the syntactic encoding of a SWD.
    - Language: It shows the language used by a SWD.
    - OWL Species: It shows the language species of a SWD written in OWL.
  - RDF Statistics: It refers to the properties summarizing node distribution of the RDF graph of a SWD.
  - Ontology annotation: It refers to the properties that describe a SWD as an ontology.
- **Relations:** They consider the explicit semantics between individual SWDs. Table I shows the different types of relations which can classified into four categories:
  - TM/IN captures term reference relations between two SWDs.
  - IM shows that an ontology imports another ontology.
  - EX shows that an ontology extends another.
  - PV shows that an ontology is a prior version of another.

- **Analytical Metadata**: It contains the SWO/SWDB classification and SWD ranking. The proposed indexing and ranking technique will discuss in the following subsection.

Table 1: The Types of the relationships among SWDs.

Type	**Classes and Properties**
TM/IN	owl:termRef, daml:termRef
IM	owl:imports, daml:imports
EX	rdfs:subClassOff, rdfs:subPropertyOf, owl:disjointWith, owl:equivalentClass, owl:equivalentProperty, owl:complementOf, owl:inverseOf, owl:intersectionOf, owl:unionOf, daml:sameClassAs, daml:samePropertyAs, daml:inverseOf, daml:disjoinWith, daml:complementOf, daml:unionOf, daml:disjoinUnionOf, daml:ntersectionOf
PV	owl:priorVersion, owl:DeprecatedProperty, owl:DeprecatedClass, owl:backwardCompatibleWith, owl:incompatibleWith

3.4 Indexing and Ranking Stage

We propose a general approach for Semantic ranking to provide high quality, high recall search in databases and on the Web. A hybrid page ranking technique is proposed which integrate the strength of both ObjectRank [19] which is calculated offline and the Hits [20] search which is run online. Therefore, our hybrid approach is using a number of relatively small subsets of the data graph in such a way that any keyword query can be answered by high ranked documents with only one of the sub-graphs. Our proposed approach tries to find the trade-off between query execution time and quality of the results.

Our technique is divided into two portions: pre-processing and query-time stages. At pre-processing stage, we will apply a modified version of the ObjectRank technique and HITS technique will be applied at the query time. We proposed a combination of these two techniques to avoid the pitfalls of each technique. We also want to benefit from the advantages of both of them. We will discuss these pitfalls and advantages in the following sub-sections.

ObjectRank inspired by the idea of PageRank [21] technique. These algorithms that use PageRank require a query-time PageRank-style iterative computation over the full graph. This computation is too expensive for large graphs, and not feasible at query time, as it requires multiple iterations over all nodes and links of the entire database graph.

On the other hand, one advantage of the HITS algorithm is its dual rankings. HITS presents two ranked lists to the user: one with the most authoritative documents related to the query and the other with the most "hubby" documents. Authoritative pages relevant to the initial query should not only have large in-degree; since they are all authorities on a common topic, there should also be considerable overlap in the sets of pages that point to them. Thus, in addition to highly authoritative pages, the researchers expect to find what could be called hub pages: these are pages that have links to multiple relevant authoritative pages. It is these hub pages that "pull together" authorities on a common topic, and allow us to throw out unrelated pages of large in-degree. A good hub is a page that points to many good authorities; a good authority is a page that is pointed to by many good hubs

To solve the problems of the ObjectRank and HITS techniques, we join these techniques as follows. First, we calculate the ranking scores for all the SWDs in our database using a modified version of the ObjectRank algorithm save them in a repository indexed with keywords. ObjectRank gives the same initial values for all nodes. In our experiment, we initialized each node with the ratio of all links that the node receives as in-links instead of giving the same initial value for all the pages. The ratio offers an enhanced initial guess with minimal overhead. In experimental evaluation, we found that this initial hypothesis reduces the number of iterations required by about one third. When a user type a query the technique will work as follow:

- We search in the repository for the most n ranked SWDs. this will reduce the time ObjectRank need to look at the entire database for a very large number of output.
- Make a sub-graph around these SWDs by adding the in-links and out-links SWDs.
- for each page in the sub-graph add only d pages from the pages that point to it (in-links), then add all pages that this page point to (out-links).
- Calculate the hub score and authority score for each page in the sub-graph.
- Output the most authoritative pages and the most hubby pages to the user.

The PageRank algorithm evaluates the Web documents according to the following equation:

$$PR(A) = (1-d) + d(PR(T_1)/C(T_1) + PR(T_2)/C(T_2) + ... + PR(T_n)/C(T_n))$$

where A is a document. T_1; T_2; .... ; T_n are Web documents that link to A; $C(T_i)$ is the total out-links of T_i; and d is a damping factor, which is typically set to 0.85. This equation captures the probability that a user will arrive at a given page either by directly addressing it, or by following one of the links pointing to it.

Unfortunately, this random model is not appropriate for the Semantic Web. Because there is different types of link between SWDs. Therefore, this leads to a non-uniform probability of following a particular outgoing link, So

ObjectRank uses a model which accounts for the various types of links that can exist between SWDs.

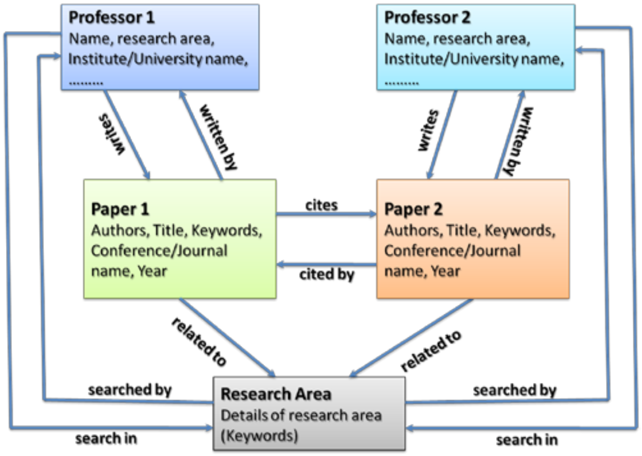

Professor 1
Name, research area, Institute/University name, ...
Professor 2
Name, research area, Institute/University name, ...
writes
written by
writes
written by
Paper 1
Authors, Title, Keywords, Conference/Journal name, Year
cites
cited by
Paper 2
Authors, Title, Keywords, Conference/Journal name, Year
related to
related to
searched by
searched by
Research Area
Details of research area (Keywords)
search in
search in

Fig. 3 An example of a subset of the ObjectRank graph.

Consequently, ObjectRank is developed as a system to perform authority-based keyword search on databases which is inspired by PageRank. Fig. 3 shows an example of a subset of ObjectRank graph. The ObjectRank algorithm applies authority-based ranking to keyword search in databases modeled as labeled graphs. Conceptually, authority originates at the nodes (objects) containing the keywords and flows to objects according to their semantic connections. Each node is ranked according to its authority with respect to the particular keywords. One can adjust the weight of global importance, the weight of each keyword of the query, the importance of a result actually containing the keywords versus being referenced by nodes containing them, and the volume of authority flow via each type of semantic connection. This algorithm as we can see is divided into two parts the preprocessing time and this what we are concerned about, and we will modify the query time stage of the ObjectRank technique.

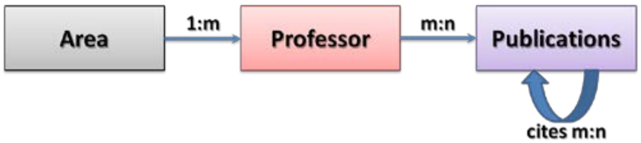


Fig. 4 The ObjectRank schema graph.

They view a database as a labeled graph, which is a model that easily captures both relational and XML databases. The data graph D(V_D, E_D) is a labeled directed graph where every node v has a label λ(v) and a set of keywords. Each node represents an object of the database and may have a sub-structure. Without loss of generality, ObjectRank assumes that each node has a tuple of attribute name/attribute value pairs. One may assume richer semantics by including the metadata of a node in the set of keywords. For example, the metadata "Forum", " Year", "Location" could be included in the keywords of a node.
Each edge e from u to v is labeled with its role λ(e). For simplicity, we assume that there are no parallel edges and we will often denote an edge e from u to v as $u \to v$.
Fig. 4 shows the schema graph which is generated from Fig. 3. The schema graph G(V_G, E_G) is a directed graph that describes the structure of D. Every node has an associated label. Each edge is labeled with a role. We say that a data graph D(V_D, E_D) conforms to a schema graph G(V_G, E_G) if there is a unique assignment μ such that:

1. For every node $v \in V_D$ there is a node $\mu(v) \in V_G$ such that $\lambda(v) = \lambda(\mu(v))$;
2. For every edge $e \in E_D$ from node u to node v there is an edge $\mu(e) \in E_G$ that goes from $\mu(u)$ to $\mu(v)$ and $\lambda(e) = \lambda(\mu(e))$.

From the schema graph G(V_G, E_G), we create the authority transfer schema graph \$$G^A(V_G, E^A)$ to reflect the authority flow through the edges of the graph. This may be either a trial and error process, until we are satisfied with the quality of the results, or a domain expert's task. In particular, for each edge $e_G = u \to v$ of E_G, two authority transfer edges, $e_G^f = (u \to v)$ and $e_G^b = (v \to u)$ are created. The two edges carry the label of the schema graph edge and, in addition, each one is annotated with a (potentially different) authority transfer rate - $\alpha(e_G^f)$ and $\alpha(e_G^b)$ correspondingly. We say that a data graph conforms to an authority transfer schema graph if it conforms to the corresponding schema graph. (Notice that the authority transfer schema graph has all the information of the original schema graph.) Fig. 5 shows the authority transfer schema graph that corresponds to the schema graph in Fig. 4 (the edge labels are omitted). The motivation for defining two edges for each edge of the schema graph is that authority potentially flows in both directions and not only in the direction that appears in the schema. For example, a paper passes its authority to its authors and vice versa. Notice however, that the authority flow in each direction (defined by the authority transfer rate) may not be the same. For example, a paper that is cited by important papers is clearly important but citing important papers does not make a paper important.

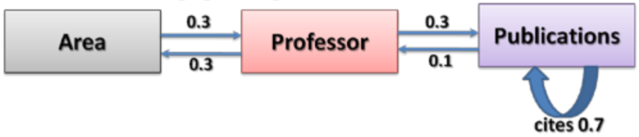


Fig . 5 The ObjectRank authority transfer schema graph.

Given a data graph $D(V_D, E_D)$ that conforms to an authority transfer schema graph $G^A(V_G, E^A)$, ObjectRank derives an authority transfer data graph $D^A(V_D, E_D^A)$ as follows.

For every edge $e = (u \to v) \in E_D$ the authority transfer data graph has two edges $e^f = (u \to v)$ and $e^b = (v \to u)$. The edges e^f and e^b are annotated with authority transfer rates $\alpha(e^f)$ and $\alpha(e^b)$. Assuming that e^f is of type e_G^f, then

$$\alpha(e^f) = \begin{cases} \dfrac{\alpha(e_G^f)}{OutDeg(u, e_G^f)} & if \quad OutDeg(u, e_G^f) > 0 \\ 0 & if \quad OutDeg(u, e_G^f) = 0 \end{cases}$$

where $OutDeg(u, e_G^f)$ is the number of outgoing edges from u, of type e_G^f. The authority transfer rate $\alpha(e^b)$ is defined similarly. Fig. 6 illustrates the authority transfer data graph that corresponds to the data graph of Fig. 3 and the authority schema transfer graph of Fig. 4. Notice that the sum of authority transfer rates of the outgoing edges of a node u of type $\mu(u)$ may be less than the sum of authority transfer rates of the outgoing edges of $\mu(u)$ in the authority transfer schema graph, if u does not have all types of outgoing edges.

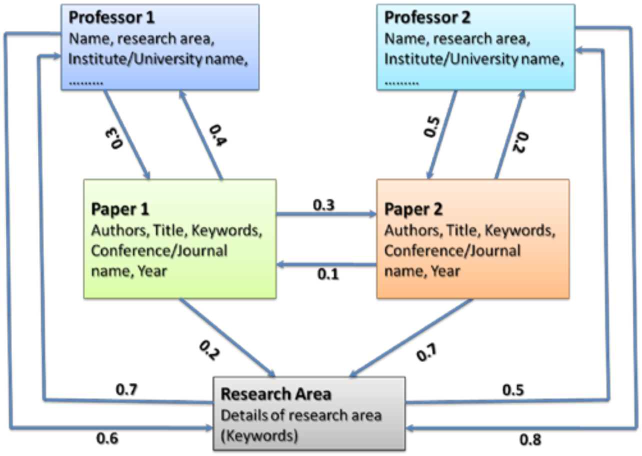

Professor 1
Name, research area, Institute/University name, ...
Professor 2
Name, research area, Institute/University name, ...
Paper 1
Authors, Title, Keywords, Conference/Journal name, Year
Paper 2
Authors, Title, Keywords, Conference/Journal name, Year
Research Area
Details of research area (Keywords)
0.3
0.4
0.5
0.2
0.3
0.1
0.2
0.7
0.7
0.5
0.6
0.8

Fig. 6 Authority transfer data graph.

Then the total score of a page will be

$$r_G(v) = (1-d) + d \sum A_{ij}$$

where $A_{ij} = \alpha(e)$ if there is an edge $e = v_j \to v_i$ in E_D^A and 0 otherwise, the damping factor d determines the portion of ObjectRank that an object transfers to its neighbors as opposed to keeping to itself. It was first introduced in the original PageRank technique [21], where it was used to ensure convergence in the case of PageRank sinks. The value for d used by PageRank is 0.85.

On the other hand, HITS is applicable in Semantic Web. The Semantic Web graph can be described by an adjacency matrix. For a network graph matrix M the well known authority ranking methods like HITS can be applied. HITS defines the authority ranking problem through mutual reinforcement between so-called hub and authority scores of graph nodes. The authority (relevance) score of each node is defined as the sum of hub scores of its predecessors. Analogously, the hub (connectivity) score of each node is defined as a sum of the authority scores of its successors.

The HITS team makes use of the relationship between hubs and authorities via an iterative algorithm works as follow: with each page A, they associate a non-negative authority weight x_{hpi} and a non-negative hub weight y_{hpi}.

They maintain the invariant that the weights of each type are normalized so their squares sum to 1:(Here also we need to take in our concern the type of the relations that exist between the SWDs we need to add this to the equations)

$$\sum_{A \in S_\sigma} (x^{\langle A \rangle})^2 = 1$$

$$\sum_{A \in S_\sigma} (y^{\langle A \rangle})^2 = 1$$

They view the pages with larger x and y-values as being "better" authorities and hubs respectively. If A points to many pages with large x-values, then it should receive a large y-value; and if A is pointed to by many pages with large y-values, then it should receive a large x-value. This motivates the definition of two operations on the weights, which denote by I and O. The I operation updates x-weights as follows:

$$x^{\langle A \rangle} \leftarrow \sum_{q:(q,A) \in E} y^{\langle q \rangle}$$

The O operation updates the y-weights as follows:

$$y^{\langle A \rangle} \leftarrow \sum_{q:(q,A) \in E} x^{\langle q \rangle}$$

Thus I and O are the basic means by which hubs and authorities reinforce one another. Therefore, to find the desired "equilibrium" values for the weights, one can apply the I and O operations in an alternating fashion, and see whether a fixed point is reached.

3.5 Metadata and Ontologies Repository

The metadata and ontologies repository is created to store the processed data for each SWD. The stored data can be used to derive analytical reports, such as classification of

SWOs and SWDBs, rank of SWDs, and the IR index of SWDs.

3.6 User Interface

Every time a user submits a query, the proposed system analyzes it and tries to identify the ontological elements which are stored in the metadata and ontologies repository. Then, it is able to suggest to the user the potential meanings of his query that it recognized. The user is therefore presented with both the results of the syntactic search and the available meanings extracted from the query. This can help him to refine his request, disambiguating among its the possible acceptations. When a user query is re-conducted to a specific meaning, the proposed system is able not only to look up resources semantically related to that meaning, but also to seek other concepts that could be of interest for the user. This is possible because the system can navigate across the sub-graph of interconnected elements of the domain ontology to generate the corresponding hubs and authorities.

4. Implementation And Results

For our experiments, we implemented our system in Java. The experiments were performed on a single PC with an Intel 1.73 GHZ Duo processor with 3GB RAM. We run an experiment to measure the effect of the total size of the sub-graph on the quality of the result. The total size of the sub-graph depends on two parameters. The first parameter is n which represents the number of pages that the sub-graph should start with. The second parameter is d which presents the number of pages a single page can bring into the sub-graph from the pages that pointing to it.

For our experiment we generate a comprehensive set of sub-graphs with 24 combinations of n and d. for each combination we measure the performance of our rank, i.e. the query time an quality of two lists.

Fig. 7 shows the effect of d on sub-graph construction time. Bigger d implies that more time to construct the sub-graph. Therefore, the quality of our rank algorithm is strongly affected by d. Thus, one has to strike balance between the quality of results and the time needed to construct the sub-graph.

5. Conclusion

Search engines are becoming such a powerful tools not only to find textual resources but also to analyze the contents of the document to get precise search result. Therefore, syntactic techniques are used to extract ontologies and metadata from the SWDs to calculate an accurate classification of the processed documents. The lexical and conceptual characteristics of a domain in an ontology are captured to prove that Semantic Web technologies provide real benefits to end users in terms of an easier and more effective access to information.

We developed an indexing and ranking technique which can be used in a Semantic Web search engine to facilitate the development of the Semantic Web and finding a proper ontology for the submitted search query. The entire documents are processed to extracted ontologies and find the relationships between the processed documents and the others. Therefore, our system is not based only on the extracted metadata from the document but also on extracted ontologies and the relations between documents. In other words, our main goals were to find a good measure for indexing the processed SWDs with extracting the proper ontologies, create a meaningful rank measure which reflects the importance of the processed document, and answer the user's queries efficiently.

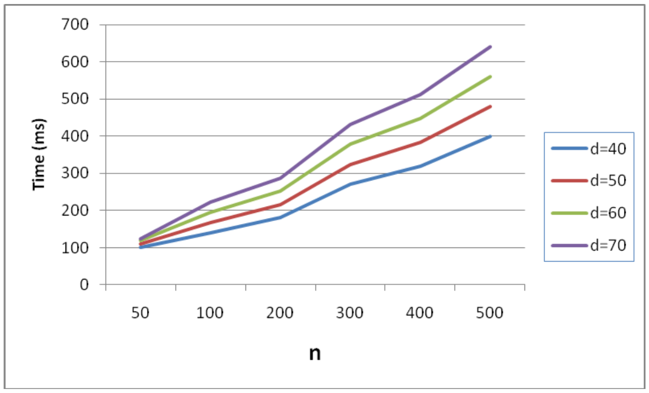


Fig. 7 The effect of n and d values on processing time.

References

[1] H. Chen, "Machine Learning for Information Retrieval: Neural Networks, Symbolic Learning, and Genetic Algorithms," Journal of the American Society for Information Science , vol. 46, no. 3, pp. 94-216, 1995.

[2] R. Baeza-Yates. Modern Information Retrieval, Addison Wesley, 1999.

[3] T. Berners-Lee, "Semantic Web Road map," http://www.w3.org/DesignIssues/Semantic.html (Last access on 2.7.2011).

[4] E. Andersen," Edging Toward the Semantic Web: Protocols, Curation, and Seeds," Ubiquity 2010, ACM, Nov. 2010.

[5] W. Yong-gui and J. Zhen, "Research on semantic Web mining," In the proceedings of the 2010 International Conference on Computer Design and Applications (ICCDA), pp. 67-70, 2010.

[6] R. Guha, R. McCool and E. Miller, "Semantic Search," In the proceedings of the 12th international conference on World Wide Web (WWW'03),pp. 700-709 , 2003.

[7] R. Guha and R. McCool, "TAP: A Semantic Web Platform," The International Journal of Computer and Telecommunications Networking - Special issue: The Semantic Web: an evolution for a revolution , vol. 42, no. 5, pp. 557-577, 2003.

[8] A. Kiryakov, B. Popov, D. Ognyanoff, D. Manov, A. Kirilov and M. Goranov, "Semantic annotation, indexing, and retrieval," Web Semantics: Science, Services and Agents on the World Wide Web , vol. 2, no. 1, pp. 49-79, 2004.

[9] H. Cunningham, D. Maynard, K. Bontcheva and V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In the proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02), 2002.

[10] U. Shah, T. Finin, A. Joshi, R . S. Cost and J. Matfield. Information retrieval on the semantic web. In the proceddings of the 11th international conference on Information and knowledge management, pp. 461-468, 2002.

[11] J. Mayfield and T. Finin. Information retrieval on the Semantic Web: Integrating inference and retrieval. In the proceedings of the SIGIR Workshop on the Semantic Web, 2003.

[12] T. Finin, J. Mayfield, C. Fink, A. Joshi and R. Scott Cost. Information Retrieval and the Semantic Web. In the proceedings of the 38th International Conference on System Sciences, 2005.

[13] I. Celino, E. Della Valle, D. Cerizza and A. Turati. Squiggle: a Semantic Search Engine for Indexing and Retrieval of Multimedia Content. In the proceedings of the 1st International Workshop on Semantic-enhanced Multimedia Presentation Systems, pp. 20-34, 2006.

[14] L. Ding, T. Finin, A. Joshi, Y. Peng, R. Pan, P. Reddivari, R. Scott Cost, V. Doshi and J. Sachs. Swoogle: A Semantic Web Search and Metadata Engine . In the proceedings of CIKM'04, 2004.

[15] T. Finin, Y. Peng, P. Reddivari, R. Scott Cost, R. Pan, J. Sachs, V. Doshi, A. Joshi and L. Ding. Swoogle: A Search and Metadata Engine for the Semantic Web . In the proceedings of the 13th ACM Conference on Information and Knowledge Management, pp. 652-659, 2004.

[16] N. Stojanovic, A. Maedche, S. Staab, R. Studer and Y. Sure. SEAL: a framework for developing SEmantic PortALs. In the proceedings of the 1st international conference on Knowledge capture, pp. 155-162, 2001.

[17] A. Yousefipour, A. G. Neiat, M. Mohsenzadeh, and M. S. Hemayati, "An ontology-based approach for ranking suggested semantic web services," In the proceedings on the 2010 6th International Conference on Advanced Information Management and Service (IMS), pp. 17-22, 2010.

[18] JENA: http://jena.sourceforge.net/ (Last Access on 30.7.2011).

[19] A. Balmin, V. Hristidis and Y. Papakonstantinou. ObjectRank: Authority-Based Keyword Search in Databases. In the proceedings of the 2006 ACM SIGMOD international conference on Management of data, pp. 796-798, 2006.

[20] H. Marmanis and D. Babenko. Algorithms of the Intelligent Web, Manning Publications, 2009.

[21] A. N. Langville and C . D. Meyer, "Deeper inside PageRank," Internet Mathematics , vol. 1, no. 3, pp. 335-380, 2003.

Ahmed Tolba got his PhD degree in Electrical Engineering from Wuppertal University, Germany, in 1988, on Computer Vision. He is working as a Dean of the Faculty of Computer Studies at Arab Open University in Kuwait. He is also a professor in department of Computer Science, Faculty of Computers and Information, Mansoura University, Egypt. He published more than 100 previewed papers in international journals and conferences. He is interested in Artificial Intelligence, natural language processing, computer vision, and E-learning.

Nabila Eladawi got her B.Sc. in information systems from faculty of computers and information , Mansoura University, Egypt, in 2002. She is working as a demonstrator in the department of information systems, faculty of computers and information, Mansoura University, Egypt. She is interested in Semantic Web, Search engines, and natural language processing.

Mohammed Elmogy got his PhD degree in computer science from Hamburg University, Germany, in 2010, on Robotics. He is working as an assistant professor in information systems department, faculty of computers and information, Mansoura University, Egypt. He is interested in Robotics, Artificial Intelligence, Semantic Web, and Computer Vision.

Solving touristic trip planning problem by using taboo search approach

Kadri Sylejmani[1, 2] and Agni Dika[1]

[1] **Department of Computer Engineering, Faculty of Electrical and Computer Engineering, University of Prishtina**
Prishtina, 10000, Kosovo

[2] **Faculty of Informatics, Vienna University of Technology**
Vienna, A-1040, Austria

Abstract

In this paper, we introduce an algorithm that automatically plans a touristic trip by considering some hard and soft constrains. Opening and closing hours of POIs (Points of Interest), trip duration and trip allocated budget represent the hard constraints, while the satisfaction factors of the POIs and travelling distance in the trip are considered as soft constraints. We use the soft constraints to evaluate the generated solution of the algorithm. The algorithm is developed by utilizing the taboo search method as a meta heuristic. The operators of Swap, Insert and Delete are used to explore the search space. The Swap and Insert operator are used in each iteration of the algorithm loop, while the Delete operator is used whenever the algorithm tends to enter in an endless cycle. The algorithm is developed by using Java programming language, while the data repositories are created in the XML format. The algorithm is tested with 40 instances of POIs of the city of Vienna. Various entry parameters of the algorithm are used to test its performance. The results gained are discussed and compared in respect to the optimal solution.

Keywords: *point of interest, optimization, planning, Swap, Insert, Delete.*

1. Introduction

Tourists that visit one city or region during a trip of a limited time, find it impossible to visit all POIs that exist in that particular area. Thus, they have to select some POIs that they consider as more interesting and worthy for them. Doing the plan of visit that includes most interesting POIs to visit, for the available time, is usually a complex task. In such situations, it would be helpful for the tourist to have a system that runs on a hand held device, which would enable him to automatically plan the touristic trip. In general, systems like that tend to fulfill as much as possible the satisfaction of tourist by making a personalized trip plan. Usually, the constraints considered by planning systems under discussion are: geographical locations of POIs and their opening and closing hours, personal score of each POI for the tourist, duration of the trip, etc. In order to produce a trip that fits all/most of these constraints, a heuristic that tends to find an optimized trip needs to be introduced.

The simplest problem in trip planning can be compared to the Orienteering Problem (OP) [1], where a number of *n* locations are given, each of them having a score *s*. The goal is to have a single tour trip that includes as many points as possible, so the satisfaction factor of the trip is maximized. The Team Orienteering Problem (TOP) is an extended form of OP, which generalizes the problem for multiple tours [2]. Further, the Team Orienteering Problem with Time Windows (TOPTW) is an advanced version of TOP, where each location is associated with a time window that represents the timings when the visit could be realized [3]. In TOPTW the goal is to determine *m* routes, each limited by T_{max}, that maximizes the total collected score. In fact, the TOPTW is a simplified version of the Tourist Trip Design Problem (TTDP) [4].

An additional feature associated to these planning systems is that they have to plan the trip in real time, so that they can respond to changing user requests and preferences as well as unexpected events. In order to achieve optimal values, for such hard planning problems, for the execution time of around tens of seconds, we need to use a meta-heuristic to solve the problem. For instance, when more time than planned, is spent to visit a particular POI, tourists would like to be able to generate an updated trip plan in real time (possibly not lasting more than some tens of seconds).

The main contribution of this paper is introduction of an algorithm that finds optimal solutions for trip planning problem, in the execution time of some tens of seconds (varying from the particular details of the trip). The goal is to find a solid solution (not the best possible) in less than 10 seconds. This can be achieved by having a fast evaluation process of the candidate solutions, and by utilizing flexible operators to explore the search space. In this paper, we use taboo search heuristic as a guiding method for the search process. The taboo search heuristic uses the memory lists to save the prior search information, which, afterwards, is used to guide the search process towards finding the global optimal solution.

In the next section a literature review is presented and in Section 3 a problem definition is introduced. In Section 4 the detailed description of the algorithm is given, while in Section 5 the experimental results are shown. Discussions, conclusions and future work are elaborated in Section 6 and 7 respectively.

2. Literature review

There are many algorithms that deal with OP, TOP or TOPTW that are discussed in the literature. These algorithms could be also applied for the purpose of trip planning. A taboo search heuristic that effectively solves the TOP is presented in [5]. Roughly, this heuristic finds the solution by iteratively repeating the steps of initialization, solution improvement and evaluation. It utilizes a number of input parameters, which are used to fine tuning the performance of the algorithm.

Another algorithm, which is based on Guided Local Search (GLS) method [7] and can solve the TOP problem as well, is elaborated in [6]. In consecutive iterations of the algorithm, GLS method does the penalization of specific unwanted solutions. The penalization operator decreases the value of evaluation function for the specific solutions. This enables the algorithm to escape from getting stuck in the local optimum and carry on with further searches in different regions of search space.

An algorithm that is based on Variable Neighborhood Search (VNS) method [8], is presented in [9]. VNS systematically searches for a better solution, by changing the procedure of neighborhood creation. This changes the search direction either towards the optimal solution (local search) or towards the opposite direction (shaking of search process). This method, in its basic version, has the advantage of not requiring too many input parameters, while being able to produce solutions of high quality.

In [10] a simple algorithm that belongs to the family of Iterated Local Search (ILS) [11] is presented. Based on the experimental results, this algorithm is able of finding good solutions when applied for the data sets known in the literature. In a combined way, the operators of *Insertion* and *Shake* are applied. Instead of finding a number of random ILS solutions, the algorithm does build a sequence of solutions, obtained with the local search method. In this algorithm, it is important to have a balance between the number of iterations of algorithm execution and the frequency of *Shake* operator utilization.

Greedy Randomized Adaptive Search Procedure (GRASP) [12] is first used to solve the TOP in [13]. In general, GRASP method is executed for a pre specified number of iterations, where initially the procedure for solution construction is executed, followed by a procedure for local search. The behavior of the procedure for solution construction is controlled by the so called parameter “Greediness”, which represents the quotient between the Greediness and Randomness of the algorithm. This quotient shows how much the algorithm uses Greediness approach compared to Randomness approach, or vice versa. Different iterations of the algorithm are independent from each other, which mean that they return independent results.

In a Tourist Information System (TIS) presented in [14], authors use a trip planning algorithm that is developed using genetic algorithms. The path for visiting the selected POIs is created in two separate steps. The first step does the calculation of the shortest path between each and every POI, by using A* algorithm. The second step decides about the order of visits to particular POIs. In this case, the genetic algorithms are used to create a list of candidate solutions. Furthermore, regardless of the execution time of the algorithm, an approximated solution is always returned. By using genetic algorithms, the algorithm under discussion, is able to propose multiple paths to the tourists. This flexible feature will allow them to select one of the proposed routes.

3. Formulation as mathematical problem

The proposed algorithm lies on the field of optimizations of touristic tours, where a number of constraints are considered for planning and optimization of the tour. The goal is to plan a multiple day trip that will serve the tourist to visit a number of touristic sites/POI (Point Of Interests). This problem can be considered as a version of Orienteering Problem with Time Windows (OPTW). A set of *n* locations is given, where each of them *(i=1,…, n)* is associated with a satisfaction value S_i ,an entrance fee f_i, a typical visit duration *ti* an opening *(Oi)* and closing *(Ci)* hour. Usually the trip has several tours, with breaks in between (night time and mid daybreaks). Each tour is limited to a maximal period of time *Tmax* and it starts at a particular fixed point and ends at another fixed point. In general, the starting / ending time and tour duration are variable for each particular day. Mostly, the starting and ending point are the same for a tour of a single day (e.g. the tour starts and finishes at the hotel). The time *tij* needed to travel from location *i* to *j*, and vice versa, is known for all locations. In general, not all locations can be visited during the trip, since the duration of the trip is limited to m tours and the tours themselves are limited to T_{max}. Each location can be visited at most once. The visit is associated with a maximum budget B_{max}, which should not be exceeded throughout the entire trip.

No waiting times are considered at the POIs, meaning that the tourist will not have to wait any time prior to the realization of the visit to the POIs. This determines an additional constraint that makes sure that the timings of visits v_i are scheduled only when POIs are open.

The aim is to find a trip with m tours that includes as many available POIs as possible, by ensuring that the trip remains under budget and also taking in to account the total satisfaction factor and travel time of the trip. The intention is to have a higher satisfaction factor and shorter travel time. The trip budget and duration is considered as hard constraint, while the constraints of satisfaction factor and travel time are taken as soft constraints, and as such take part in the evaluation of the proposed solution.

The constraints of satisfaction factor and travel distance are non proportional between themselves. For instance, if there are more POIs in the trip, the satisfaction factor will be higher, while the travel time will be higher too, which conflicts with the travel time aspiration constraint. On the other hand, in order to degrease the travel time, it is needed to have less visits in the trip, which would minimize the satisfaction factor, which again opposes the intention to get a maximal satisfaction factor for the trip. Hence, defining an evaluation function of the trip that enables finding the optimal value, of both satisfaction factor and travel time constraints, is needed.

Based on the facts elaborated above, the running planning problem can be defined with following mathematical expressions:

$$Max\{\textstyle\sum_{i=1}^{n}(S_i * x_i)\} \qquad (1)$$

$$\textstyle\sum_{i=1}^{n}(B_i * x_i) \le Bmax \qquad (2)$$

Where:

$$x_i = \begin{cases} 1, & if\ point\ \boldsymbol{i}\ is\ visited\ during\ the\ trip \\ 0, & if\ point\ \boldsymbol{i}\ is\ not\ visited\ during\ the\ trip \end{cases}$$

n – Number of available POIs for visit
S_i – Satisfaction factor of point ***i***
B_i- Entry fee of point ***i***

$$Min\left\{\left[\sum_{i=1}^{n}\left(\sum_{\substack{j=1\\ j\neq i}}^{n} t_{ij} * y_{ij}\right)\right] + \sum_{i=1}^{n}(t_{si}\,u_i + t_{ie}v_i)\right\} \qquad (3)$$

Where:

$$y_{ij} = \begin{cases} 1, & if\ a\ visit\ to\ \boldsymbol{i}\ is\ followed\ by\ a\ visit\ to\ \boldsymbol{j} \\ 0, & if\ a\ visit\ to\ \boldsymbol{i}\ is\ not\ followed\ by\ a\ visit\ to\ j \end{cases}$$

$$u_i = \begin{cases} 1, & if\ visit\ \boldsymbol{i}\ is\ first\ visit\ in\ the\ tour \\ 0, & if\ visit\ \boldsymbol{i}\ is\ not\ first\ visit\ in\ the\ tour \end{cases}$$

$$v_i = \begin{cases} 1, & if\ visit\ \boldsymbol{i}\ is\ last\ visit\ in\ the\ tour \\ 0, & if\ visit\ \boldsymbol{i}\ is\ not\ last\ visit\ in\ the\ tour \end{cases}$$

t_{ij} – travel time from point ***i*** to ***j***
t_{si} – travel time from start point to point ***i***
t_{ei} – travel time from point ***i*** to end point

$$\textstyle\sum_{i=1}^{m} z_{ij} \le 1 \qquad \text{(j=1, …, n)} \qquad (4)$$

Where:

$$z_{ij} = \begin{cases} 1, & if\ visit\ \boldsymbol{i}\ in\ trip\ is\ point\ \boldsymbol{j} \\ 0, & if\ visit\ \boldsymbol{i}\ in\ trip\ is\ not\ point\ \boldsymbol{j} \end{cases}$$

m – Number of POIs visited during the entire trip

$$o_i \le v_i\ \ \&\ \ v_i + t_i \le\ c_{i,}\ \ i = 1, \dots, m \qquad (5)$$

Expression (1) defines the intended maximal satisfaction factor of the trip, while expression (2) ensures that trip is equal or lower than the budget allocated for the trip. Formula (3) expresses the minimal travel time aspiration, by considering the travel times between visited points themselves and also between them and the starting/ending points. Expression (4) makes sure that a particular point is visited at most one time, while expression (5) makes the trip feasible only when all the points of interests are open on their scheduled time.

4. Description of the algorithm

Overview

Trip planning is done based on the entry data that describe the trip. The entry data sets are categorized in three different kinds:

- Data that describe the trip, such as: start/end date of trip, allocated budget, accommodation location, and number of tours to be taken during the trip, coefficient of weight of satisfaction factor and travel time, tourist preferences for categories and types of POIs and an additional entry parameter that specifies one of the two possible regimes of work of the algorithm.
- Data that describe the POIs, such as: name of the POI, typical visit duration, location, entry fee, working hours, type and category of POI.
- Data about travel distances between each and every POI that is available. The travel distances are expressed in unit of minutes.

As seen in the figure 1, the trip planning algorithm consists of two separate modules. One of them deals with calculation of personal score for the POIs, while the other one does the actual planning of the trip. In this paper, calculation of personal score (satisfaction factor), which is the process that is known as matchmaking between tourist preferences and POIs, is done by utilizing a simple algorithm introduced by Souffriau & Maervoet et al [15]. The value range of satisfaction factor produced by this matchmaking algorithm is between 0 and 48. Since, in our case we assume that the range of values for satisfaction factors of POIs is between 0 and 100, we have used a transformation function to convert the range of values from *[0 – 48]* to the range *[0 – 100]*.

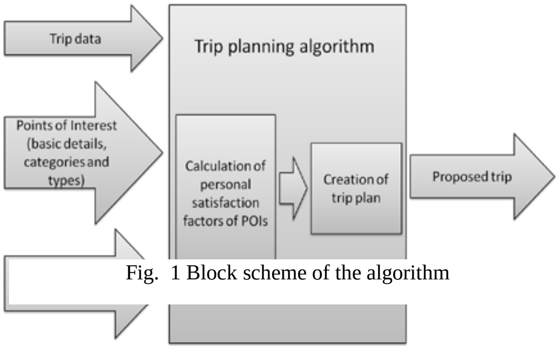

Trip data
Points of Interest (basic details, categories and types)
Trip planning algorithm
Calculation of personal satisfaction factors of POIs
Creation of trip plan
Proposed trip

Fig. 1 Block scheme of the algorithm

$$Satisfaction\ factor_{(Max=100)} = \frac{100 * [Satisfaction\ factor_{(Max=48)}]}{48}$$

Module for trip planning

The process of trip planning creates an itinerary that consists of predefined number touristic tours to be taken during the trip period. The optimization of the trip planning is done by utilizing the taboo search heuristic. In our case, the taboo search heuristic uses the operators of Swapping and Insertion for exploring the search space. The Delete operator is used in some iteration to escape the local optimum. The taboo search heuristic is known for its process of memorizing previous search information, which facilitates the escape from local optimum by changing the search direction. In our example, the planning module can be customized by nine different entry parameters, as shown in the pseudo code given below.

```
Algorithm Main(TLS, MT, MTWI, MBTNTS, ACI, PC, FMH, DN,
        FTWMV)
begin
    Operators = {Swap, Insert};
    Initialize taboo memories;
    Create initial solution $S_c$;
    Evaluate $S_c$;
    $S_b = S_c$;
    iterationNumber = 0; IterationsWithoutImprovement = 0;
    while (iterationNumber <= MT) do
        Divert=(iterationsWithoutImprovement % DN) == 0;
        for each operator in Operators do
            Generate neighbourhood of $S_c$ by using current operator;
            Find best non taboo and taboo neighbour of $S_c$ (Divert);
            if IterationsWithoutImprovement greater than ACI then
                AspirationCriteria=best taboo nighbor >
                best solution found so far;
            else
                AspirationCriteria= best taboo nighbor –
                best non taboo nighbor > MBTNTS;
            end
            if there is a feasible non taboo / taboo neighbour then
                if AspirationCriteria is fulfilled then
                    $S_c$ =Best taboo neighbour;
                else
                    $S_c$ =Best non taboo neighbour;
                end
                if operator is Swap then
                    acceptanceCriteria = $S_c$ better than $S_b$;
                else
                    if FTWMV then
                        acceptanceCriteria= $S_c$ better than $S_b$ or
                        number of visits in $S_c$ >number of visits in $S_b$ ;
                    else
                        acceptanceCriteria= $S_c$ better than $S_b$;
                    end
                end
                if acceptanceCriteria is fulfilled than
                    $S_b = S_c$;
                end
            else
                Delete a visit from trip;
            end
        next;
        if there is improvement in current iteration then
            IterationsWithoutImprovement=0;
        else
            IterationsWithoutImprovement +1;
        end
        if IterationsWithoutImprovement equals MTWI then
            Exit loop;
        end
        iterationNumber +1;
    end
end
Return $S_b$;
```

In the Table 1 we present the description for the entry parameters of the algorithm.

Table 1: Algorithm parameter description

Parameter	***Abbrev.***	***Description***
Taboo List Size	TLS	Specifies the number of iterations that a move will remain taboo. E.g. *TLS=5* indicates that swapping between point ***i*** and ***j*** cannot be performed in next five iterations.
Max Tries	MT	Indicates the total number of iterations that the algorithm will run.

Max Tries Without Improvement	MTWI	Specifies the total number of iterations that the algorithm will run without any further improvement.
Margin Between Taboo And Non Taboo Solution	MBTNTS	Indicates how much better a taboo solution should be, compared to a non taboo solution, such that it would fulfill the aspiration criteria.
Aspiration Criteria Iterations	ACI	Defines the number of iterations without improvement, which will utilize the version of aspiration criteria with margin between taboo and non taboo solution. After passing the number of iterations, indicated by ACI, the aspiration criteria is calculated in its usual form (accepting a taboo solution only if it is better than the best solution found that far).
Penalty Coefficient	PC	Takes a value between 0 and 1, which is used to penalize frequent moves that have occurred during the search process.
Frequency Memory Horizon	FMH	Determines the number of iterations after which the frequency based memory will be reset.
Diversification Number	DN	Specifies how often the search process will be diversified. Every DN iterations the diversification process will take place.
Find Trip With Maximum Visits	FTWMV	It is a logical parameter that defines one of the two possible regimes of work of the algorithm. If its vale is *True*, the algorithm will try to find the best trip with maximal number of POIs. Conversely, if its value is *False*, the algorithm will focus only in finding the best evaluated trip, even though the resulting trip may not have the maximal number of POIs.

The algorithm uses two operators for exploring the search space, which are shown in the initial part of the pseudo code. The Swap operator does the swapping of POIs that are on trip with POIs that are currently out of the trip. Insertion of new POIs into the trip is made by Insert operator. The so called Taboo Memories are used to save information about the recency and frequency of swaping and inserting individual POIs. These memories will enable the search process to avoid getting stuck in the local optimum and also direct the search process in the new regions of search space (that far not explored). Before the algorithm starts looping, an initial solution is created, which is than evaluated and accepted as best current solution. In general, the initial solution is created by randomly inserting new POIs, until there is no left space.

The algorithm will be iteratively executed by MT iterations. The Boolean variable *Divert* will be calculated for each iteration of the algorithm and it is used to decide whether the search diversification operator shall be applied in current iteration. Its value is *True* if division of variable *iterationsWithoutImprovement* and parameter *DN* returns an integer, otherwise its value is *False*.

Inside the main algorithm loop, another loop (named the *operator loop*) is executed for two times. In the first execution, the *operator loop* uses the Swap operator, while in the second time it uses the Insert operator. In both executions, with Swap and Insert operator, the generation of neighborhood is full, which means that all possible combinations are considered. Swap operator swaps each POI on trip with each POI out of the trip, while the Insert operator inserts each non included POI before and after each included POI.

After the process of neighborhood generation, each valid neighbor is evaluated and the best non taboo and neighbor of current iteration are selected. In case the evaluation is done for the neighborhood generated by Swap operator, for some specific iterations (exactly every DN iterations) the operator of penalizations is used.

After finding the best two solutions (one of them taboo and the other one non taboo), the algorithm checks whether the aspiration criteria is fulfilled. This algorithm, depending on the value of ACI parameters, works with two sorts of aspiration criteria. If value of variable *IterationsWithoutImprovement* is greater than value of parameter *ACI*, then the aspiration criteria is defined as: "*Best taboo neighbor must be better than best solutions found so far, so that the taboo neighbor could be accepted as actual solution*", otherwise, the aspiration criteria is defined as "*result of subtraction between taboo and non taboo solution should be greater than the value of parameter MBTNTS, so that the best taboo solution is accepted as actual solution*".

If at least one of the neighbor solutions (taboo or non taboo) represents a feasible solution, the algorithm carries on with selection of the aspiration criteria, otherwise, the operator that deletes a POI from the trip, is applied. The POI deletion is conducted randomly in one of the tours of the trip. If aspiration criteria is fulfilled, then best taboo neighbor is accepted as the actual solution, otherwise the best non taboo solution is accepted as current solution.

Next, the variable *acceptanceCriteria* is defined, which is used to determine whether the current solution could be accepted as best solution found so far. Depending on the value of logical parameter FTWMV, the variable *acceptanceCriteria* could be defined in two different ways. If its value is *False*, than the varable *acceptanceCriteria* will be set to allow a current solution to become the best solution found so far, only if it is better. Otherwise, when parameter *FTWMV* is *True*, a current solution can become the best solution if it has more visit on the trip, regardless that it may not have a greater evaluation then the best solution found so far.

Inside the algorithm loop, the number of iterations without improvement is counted. If this number reaches the value defined by parameter MTWI or the predefined number of maximum iterations MT exceeds, than the algorithm execution stops and the best found solution is returned.

Determining the legality of the neighbor

A neighbor would be legal if it fulfills the hard constraints:

- All visits in the trip are scheduled when respective POIs are open,
- The trip budget is not exceeded, and
- The length of each tour in the trip remains in the pre specified duration

The pseudo code for determining candidate feasibility is given in the following:

```
Determine legality of neighbor
begin
    legality=false;
    if  new vist is open in scheduled time do
        if neighbor cost is under budget do
            if neighbor is viable in time do
                legality= true;
            end
        end
    end
end
return legality;

Determine time viability of neighbor(changed tour)
begin
    viability=false;
    if length of changed tour is not grater than orginla tour length do
        vilabilty=true;
    end
end
Return viabilty;
```

Evaluation function

Evaluation of the candidate solution is done by considering two soft constraints, namely *the total trip satisfaction factor* and *total trip travel time*. The goal is to find an optimal trip that has the total satisfaction factor as higher as possible, while the travel time remains as low as possible. In order to realize this, we have used an evaluation/fitness function that consists of two components:

$$Evaluation\ function = w1 * [\,satisfaction\ factor_{norm}\,] + w2 * [travel\ time_{norm}\,]$$

Parameters $w1$ and $w2$ represent the weight coefficients for the particular components of the evaluation function. In order to have a proportional effect in to the evaluation function, when the value of individual components changes, we have used the normalized values of both components:

$$satisfaction\ factor_{norm} = 100 * \frac{total\ satisfaction\ factor}{maximal\ satisfaction\ factor}$$

Where:

$$total\ satisfaction\ factor = \sum_{i=1}^{n} SF_i$$

$$maximal\ satisfaction\ factor = 100 * \frac{\mathrm{MNP}}{\mathrm{SDT}} * [\mathrm{TDWB}\,]$$

SF_i–Satisfaction factor of POI with index i,
n – Number of POIs included into the trip,
MNP – Maximal Number of POIs that are aimed to be visited per day
SDT – Standard Duration of a Tour
TDWB – Trip Duration Without Breaks

Based on the practical experience, we consider that the maximal desired number of POIs to be visited during one day tour is 20, while the duration of the tour of one day is usually 8 hours. In order to have a realistic view for the maximal satisfaction factor, we have considered only the time when the tourist is supposed to be active in his trip (TDWB), by omitting the breaks that the tourist may take (e.g. such as sleeping at the hotel at night).

Since we use the approach of maximizing the value of evaluation function, mathematically, we would need to maximize the values of both its components. While for the satisfaction factor component this is right, for the travel time component it should be the opposite aim. Hence, in order to facilitate the maximization of both components and aim in minimizing the travel time, we try to maximize the complementary value of travel time, which in fact will minimize the travel time, by using:

$$travel\ time_{norm} = 100 * (1 - \frac{travel\ time}{\mathrm{TDWB}})$$

5. Experimental results

The algorithm is tested by utilizing 10 different instances of tourist profiles. In addition, we have used 40 instances of POIs of the city of Vienna. As a starting/ending point of each tour of the trip we have used a hotel in the same city. Travel distances between POIs are expressed in the unit of minute.

All calculations are made by using a PC with an Intel Core 2 processor with 2.0 GHz and the RAM memory of 2.55 GB.

In the following experiments, if not differently stated, we have used a trip with execution details as shown in table 2.

Table 2: Default data for experiments

Parameter	***Value***
Trip duration	Two tours, five hours each
Trip Budget	200 euro
Tour start time	11:00
Tour end time	16:00
Weight of satisfaction factor	70%
Weight of travel time	30%
Execution time of the algorithm	5 minutes

Our experiments aim in obtaining the optimal values for the entry parameters of the algorithm, such as: finding the optimal taboo list size, margin of aspiration criteria, frequency of applying the operator for search diversification etc. If not differently stated, the algorithm is executed 10 times for each instance, and then, the average values of the results of particular executions are taken.

Tests with various versions of initial solutions

We have tested the algorithm with three different kinds of versions of initial solutions:

1. *Random initial solution* – where POIs are randomly entered into the trip itinerary, as much as there is room in it.
2. *Initial solution with POIs sorted in ascending ord*er – where POIs are entered into the trip based on the value of satisfaction factor. The POIs that have lower satisfaction factor are prioritized for earlier insertion into the trip itinerary.
3. *Initial solution with POIs sorted in descending order* –in this case, as well as in the previous case, the POIs are entered into the trip based on the value of satisfaction factor. Conversely, in this case the POIs that have higher satisfaction factor have higher chance for earlier insertion into the trip itinerary.

By using instance 8, the algorithm is executed 10 times for each three different initial solutions. Respective results of the execution of the algorithm for each different initial solution are compared, and then the maximal value from one of the three initial solutions is recorded. The number of maximums shown in figure 2, indicate that random initial solution performs better than the other two initial solutions, because it has been better in seven executions compared to the other initial solutions. On the other hand, the ordered lists versions (both in ascending and descending order) have never resulted better than the other ones. In three executions, at least two of the three different initial solutions have produced the same evaluation of the produced solution.

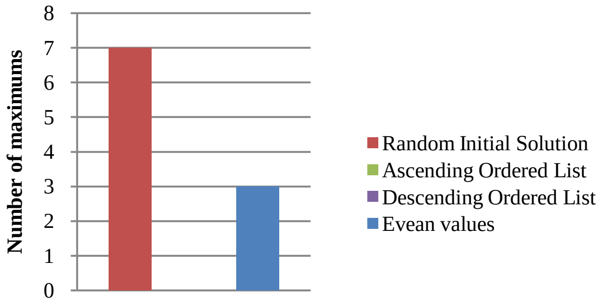


Fig. 2 Algorithm performance for different initial solutions

Variance of the algorithm result for different executions

Instance 8 is executed 10 times and the variance between different executions is shown in the following figure.

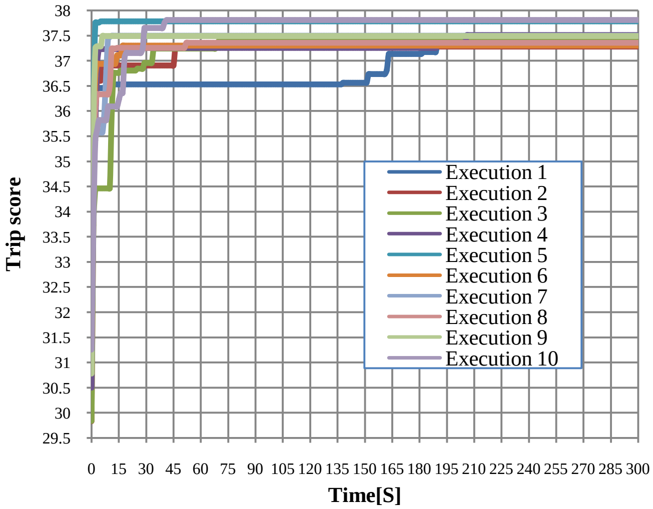


Fig. 3 Variance of the algorithm performance for different executions

It can be stated that for almost all executions, the best solution is found for approximately 40 seconds (except execution 1), and the solution with average evaluation is found for approximately 10 seconds.

Selection of taboo list size

In this experiment, all instances are executed 10 times with taboo list sizes 3, 6 and 9. Afterwards, the average values of individual executions of instances are calculated. Then the average values for individual taboo list sizes for all instances are taken. The results are shown in the below table.

Table 3 : Comparison of taboo list size

Taboo list size	Minimum	Maximum	Average	Standard deviation
3	31,376	31,599	31,444	0,072
6	31,680	31,947	**31,769**	0,084
9	31,392	31,510	31,430	0,041

Furthermore, we have also counted the number of instances for which a particular taboo list size produces better results (cf. figure 4).

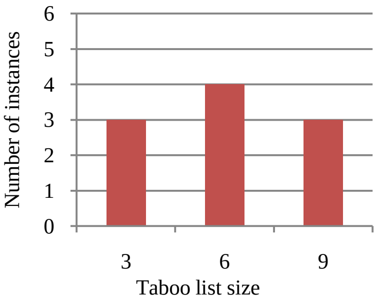


Fig. 4 Selection of taboo list size

From table 3 and figure 4, it can be conclude that taboo list size of 6 produces better results. In general, the solutions obtained by using the taboo list size of 6, are better for average 0.3 points than the solutions gained by two other taboo list sizes used in the experiment.

Diversification of search process

The diversification process ensures that the algorithm continues to search for the global optimal solution. This process is applied every *N* iterations. The experiment shows that applying the search diversification process yields to better results. Furthermore, if we apply it more often, we would gain better results.

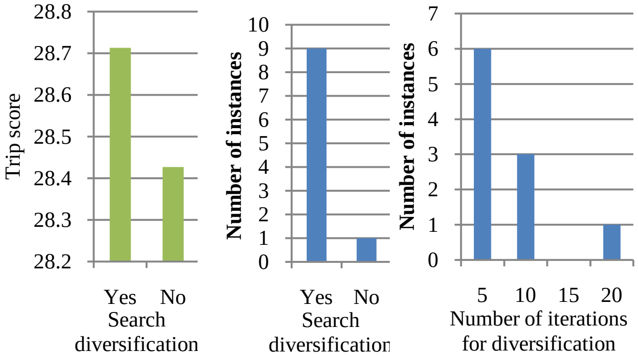


Figure 5 Advantages in applying search diversification process

Best trip versus trip with maximum POIs

The algorithm under discussion works in two different kinds of modes. The first one tries to find the highest evaluating trip, while the second one, aims in finding the best evaluating trip that has maximum number of POIs. The working mode of the algorithm is specified by the user.
The following figure, expresses a comparison between the two algorithm regimes in terms of execution time, number of POIs and evaluation.

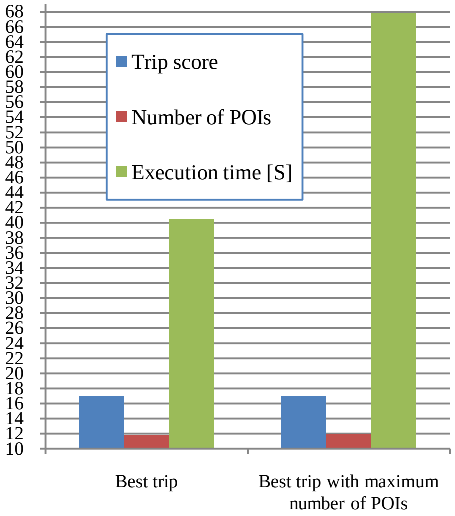


Fig. 6 Comparison of the two regimes of the algorithm

As seen in the figure, the overall trip score and number of POIs do not have a significant increase in the second mode (Best trip with maximum POIs) compared to the first mode (Best trip). Conversely, it only increases the average execution time for around 28 seconds.

Comparison of algorithm results for different tourist instances

In figure 7, we show the variation of trip score for different tourist instances. It can be noticed that for all instances, for the period of around 10 seconds, we gain solutions that evaluate near to final solutions. During this 10 second period, almost all instances are improved for about three points compared to their initial solution. After this period, no significant improvements are made (in most cases the improvement is less than one point). Hence, it can be concluded that further execution of the algorithm does not bring to significant improvement. Conversely, it will only have the negative impact of increasing the execution time of the algorithm.

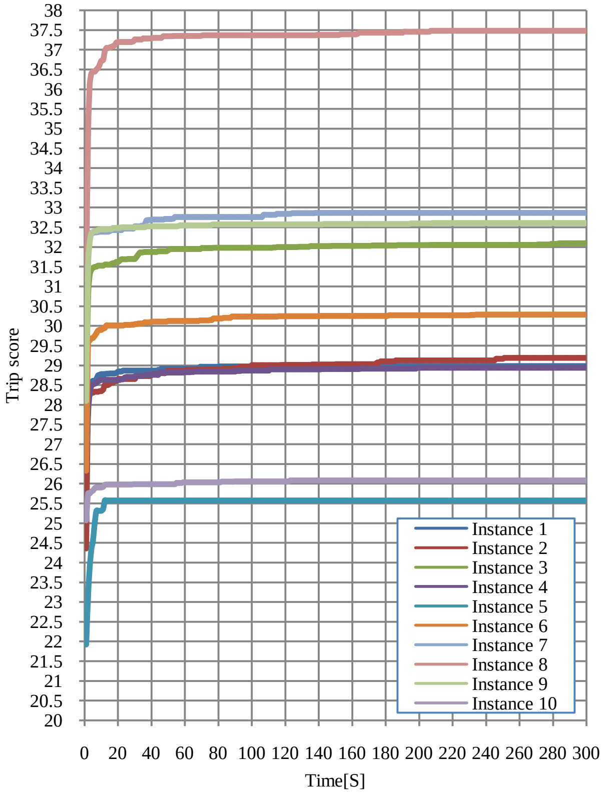


Fig. 7 Execution of the algorithm for different tourist instances

Comparison of different implementation of Swap operator

The basic implementation of Swap operator swaps each POI that is on the trip itinerary with each POI outside the trip itinerary. We have called this as "Large Swap", since the solution neighborhood is created with all possible combinations enabled by Swap operator. In addition, by using Min/Max Conflicts method, we have implemented the Swap operator in its "Small Swap" mode, where only three POIs of the current trip itinerary that have the largest travel time (travelling time from previous POI to the current POI) are considered for swapping with the POIs out of the trip itinerary. The third version of the Swap operator is implemented by using the Hill climbing method.

In figure 8, we have shown the execution of instance no. 8 with the three different versions of Swap operator. The instance no. 8 is executed 10 times with each different implementation of Swap operator, and the average values of 10 executions are presented in the figure.

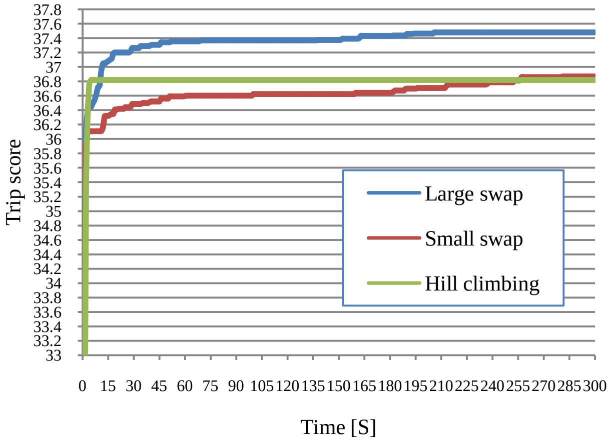


Fig. 8 Performance of the algorithm for different implementation of Swap operator

Figure 8 shows that the "Large Swap" version yields to better results, while the "Small Swap" version and the Hill Climbing method evaluate nearly to the same value. The Hill Climbing method performs faster than the other two versions (best solution is found in around 5 seconds), but quality of the solutions found by this method is worse. In addition, "Large swap" version is quicker (best solution found in approximately 210 seconds) than the "Small swap" version (best solution found in about 260 seconds).

Comparison of different implementation of Swap operator for various number of tours

In table 4, we show results gained by executing the algorithm (using instance No. 8) with different Swap operator implementation and different trip lengths.

For small number of tours (one or two tours), the "Large Swap" mode performs better than the other two implementations of Swap operator. It is noticeable that for a short duration of the trip (one or two tours), the "Large Swap" version takes only about 10 seconds more than the "Small Swap" version. On the other hand, for larger trips (three or more tours) the "Small Swap" version is quicker for about 50 seconds. Furthermore, when the trip consists of five tours, the "Small Swap" version is faster for around 130 seconds than the "Large Swap" version. Considering these results, sometimes it may be more appropriate to sacrifice a little bit in the quality of the found solution (by using the "Small Swap" version), in order generate the trip plan faster. The Hill Climbing method does not take so much time to find the final solution (in average 86 seconds), but the quality of the found solutions and the small number of POIs on those solutions, makes this method as not successful as the other two methods.

Table 4: Performance of the algorithm for different trip lengths

	Large Swap			*Small Swap*			*Hill Climbing*		
Number of tours	Trip score	Number of visits	Time [S]	Trip score	Number of visits	Time [S]	Trip score	Number of visits	Time [S]
1	37,4	4	53,6	37,35	4	52	35,15	2,6	40,2
2	35,7	8	87,9	35,41	7,9	67,9	35,47	6,7	80,7
3	34,5	11,8	127,8	33,86	11,4	81,7	34,07	10	97
4	33,4	15,5	168,1	32,76	14,3	73,3	32,86	13,2	105,6
5	32,0	18,1	226,2	31,70	17,4	96,3	31,82	15,7	109,1
Average	34,63	11,48	132,7	34,22	11,00	74,2	33,88	9,64	86,5

6. Discussions

In this paper we presented an algorithm that is used for planning the touristic trip, by considering a number of soft and hard constraints. The hard constraints consist of opening and closing hours of POIs, the trip budget and duration. The solutions generated by the algorithm are evaluated by using a fitness function that considers the overall trip satisfaction factor and tourist travel time throughout the entire trip, which in fact represent the soft constraints for the algorithm. The calculation of personal satisfaction factors for the POIs is done by using a simple algorithm introduced by [15]. The algorithm is created by using the taboo search metaheuristic, where four different kinds of initial solutions are tested. The exploration of search space is done by using the operators of Insertion, Swapping and Deletion. In order to test the performance of the algorithm, the Swap operator is implemented in three different formats. First two implementations are done by using the small and large Swap approach, respectively, while the third one is done by using the Hill Climbing method. In each iteration of the algorithm, the Insert operator tries to insert a POI in one of the available tours. The Delete operator is applied in occasional iterations, so would let the algorithm to escape from getting stack in an endless loop.

Algorithm performance test is done by conducting a number of experiments, which are mainly realized to obtain the optimal values of the entry parameters of the algorithm. The experiment with the initial solution shows that random initial solutions perform slightly better than the other ones. In addition, it is obvious that the variance between the results of different executions of the algorithm is less than one. The optimal number of iterations for which a solution would remain taboo is six. In general, solutions gained when using the taboo list size of six, score for 0.3 points more than when the taboo list size is three or nine.

It is evident that the utilization of search diversification process yields to better results. In the conduced experiments, we notice an average improvement of 0.3 points when diversification is applied. Furthermore, experimental results show that the more often we apply the diversification, the better results we gain. The penalty coefficient of 0.8 has a slight advantage in comparison to the other tested values.

Depending on the working mode of the algorithm, finding the best trip or the best trip with maximal POIs, will take approximately an average time of 40 or 60 seconds, respectively. The quality of found solutions in both regimes is nearly the same.

In terms of quality, the experiments with different implementation of Swap operator show that the “Large Swap” version outperforms the other two versions. The “Large Swap” version scores better than the “Small Swap” version and the Hill climbing method for 0.7 and 0.8 points, respectively. The Hill climbing method is able to find the final solutions in about 5 seconds, whereas the “Large and Small Swap” need much more time, which may be up to 200 or 250 seconds, respectively.

The experiments with different trip lengths show that for a trip of one or two tours, it may be more appropriate to use the “Large Swap” mode, since the quality of the solutions is better, whereas the execution time remains nearly the same to that of “Small Swap” mode. In addition, for larger number of tours (3 or more), it may be acceptable to sacrifice a little bit the quality of solutions, so that we could gain the final solution quicker by using the “Small Swap” mode.

Finally, based on the experimental results, the algorithm is able to produce a personal trip itinerary in margins of tens of seconds. Further, in order to meet specific requirements, the algorithm can be configured by using nine different parameters. The presented results indicate that for a reasonable time of execution, the algorithm generates a near to optimal trip plan.

7. Conclusions and future work

The main contribution of this paper is the introduction of an algorithm for touristic trip planning that is comparable to the well known problem of Team Orienteering Problem with Time Windows (TOPTW). In general, the trip planning can be done in an average time of 70 seconds. The solution evaluation is made by using a fitness function that consists of two separate components, where one of them considers the overall trip satisfaction factor and the other one the total traveling time. A such fitness function makes the trip plan more personal for the tourist and the algorithm suitable for use in personal trip planning systems. The algorithm performance is tested by using 40 instances of POIs of the city of Vienna and 10 different tourist profiles. The future work includes testing the algorithm with larger test instances. Additionally, testing the algorithm with test dates known in the literature will make it comparable to the existing similar algorithms. It may also be important to design new and more specific test instances, for example concerning the number of possible visits, number of tours, and the length of the time windows of POIs etc.

The relative long time to finding the optimal solution that mainly comes as the result of the process of verifying the legality of proposed solutions, may be a focus of research of work in the future. Furthermore, adding new planning constraints such as, context factors (weather, unexpected events, traffic jams, weekends etc.) could lead to more personalized trip plans.

In the real life, it often happens that a group of tourists go for a joint touristic trip. Hence, introduction of an algorithm that is able to plan a trip for group of tourists may be desired. The consideration of personal interests of individual tourists would be preferable. It would be ideal, if the algorithm could create a master trip (for the whole group) that in some portions of it could be spread into some sub trips, so that it would match interests of sub groups of tourist, who may have different preferences for specific POIs. The evaluation of the trip would need to be a general one, for the whole group of tourists, by considering a number of soft and hard constraints concerning the touristic trip.

References

[1] B.L. Golden, L. Levy, and R. Vohra. The orienteering problem. Naval Research, Logistics, 34:307-318, 1987.

[2] Chao, I.-M., B. L. Golden, E. A. Wasil. 1996b. The team orienteering problem, Eur. J. Oper. Res. 88(3) 464–474.

[3] Kantor, M. G., M. B. Rosenwein. 1992. The orienteering problem with time windows. J. Oper. Res. Soc. 43(6) 629–635.

[4] P. Vansteenwegen, D. Van Oudheusden. The mobile tourist guide: an OR opportunity. OR Insights 2007; 20(3):21-7.

[5] Tang, H., Miller-Hooks, E.: A Taboo search heuristic for the team orienteering problem. Comput. Oper. Res. 32, 1379 – 1407 (2005)

[6] P. Vansteenwegen, W. Souffriau, G. Vanden Berghe, D. Van Oudheusden, D.: A guided local search metaheuristic for the team orienteering problem. Eur. J. Oper. Res. 196(1), 118-127 (2008). Doi: 10.1016/j.ejor.2008.02.037

[7] Voudouris, C., Tsang, E.: Guided local search and its application to the travelling salesman problem. Eur. J. Oper. Res. 113, 469-499 (1999).

[8] Hansen, P., Mladenovic, N.: Variable neighbourhood search: Principles and applications, Eur. J. Oper. Res. 130, 449-467 (2001).

[9] P. Vansteenwegen, W. Souffriau, G. Vanden Berghe, D. Van Oudheusden, D.: Metahuristics for Trip Planning. Metaheuristics in the Service Industry, pages:15-31, 10.1007/978-3-642-00939-6_2, (2009).

[10] P. Vansteenwegen, W. Souffriau, G. Vanden Berghe, D. Van Oudheusden, D.: Iterated local search for the team orienteering problem with time windows, Journal of Computers & Operations Research, 36 (2009) 3281 -3290.

[11] H.R. Lourenço, O. Martin, and T. St¨utzle, “Iterated local search,” in Handbook of Metaheuristics, ser. International Series in Operations Research & Management Science, F. Glover and G. Kochenberger, Eds., Kluwer Academic Publishers, vol. 57, pp. 321–353, 2002.

[12]Feo, T.A., Resende, M.G.C. : A probabilistic heuristic for a computationally difficult set covering problem. Operations Research Letters, 867-71, (1989).

[13]Souffriau, W., Vansteenwegen, P., Berghe, G.V., Oudheusden, D.V. : A greedy randomised adaptive search procedure for the Team Orienteering Problem, EU/MEeting 2008 on metaheuristics for logistics and vehicle routing location, (2008)

[14] Maruyama, A., Shibata, N., Murat,a Y., Yasumoto, K., and Ito, M. (2004). A personal Tourism Navigation Sstem to Support Traveling Multiple Destinations with Time Restrictions. Proceedings of the 18th International Conference on Advanced Information Networking and Applications (AINA ‘04), IEEE (2004)

[15] Souffriau, W., Maervoet, J., Vansteenwegen, P., Berghe, G.V., Oudheusden, D.V. : A mobile tourist decision support system for small footprint devices, IWANN 2009, Part I, LNCS 5517, pp. 1248-1255, (2009)

First Author: Kadri Sylejmani, Dipl. Ing. in Computers and Telecommunication – 2004, Msc. in Computer Science – 2010; Teaching Assistant at Faculty of Electrical and Computer Engineering – Department of Computer Engineering, University of Prishtina, Kosovo; has presented several papers in scientific conferences and workshops on his field of research; his current research interest include filed of electronic tourism and problem solving in Artificial Intelligence.

Second Author: Agni Dika, PhD in Computer Science – 1989; Full professor at Faculty of Electrical and Computer Engineering – Department of Computer Engineering, University of Prishtina, Kosovo; his current research interest include computer logic design and algorithms.

Web Services Non-Functional Classification to Enhance Discovery Speed

Mamoun Mohamad Jamous[1], **Safaai Bin Deris**[2]

Department of Software Engineering, Faculty of Computer Science and Information Technologies

Universiti Teknologi Malaysia, Skudai, 81310, Johor Bahru, Malaysia

Abstract

Recently, the use and deployment of web services has dramatically increased. This is due to the easiness, interoperability, and flexibility that web services offer to the software systems, which other software structures don't support or support poorly. Web services discovery became more important and research conducted in this area became more critical. With the increasing number of published and publically available web services, speed in web service discovery process is becoming an issue which cannot be neglected. This paper proposes a generic non-functional based web services classification algorithm. Classification algorithm depends on information supplied by web service provider at the registration time. Authors have proved mathematically and experimentally the usefulness and efficiency of proposed algorithm.

Keywords: *Web services, web service discovery, web service classification.*

1. Introduction

Finding suitable web services for the end user or a service oriented system developer requires skills until now, and takes remarkable time even the number of web services is not very big. classification of web services is one methodology that can be used in order to enhance the speed of web service discovery process. Until now, and after the termination of the UDDI project [1], the biggest web services registry to the best of authors knowledge is Seekda [2] with 28 thousand registered web services. Another web service registry is BioCatalogue [3] with two thousand registered web services specialized in bioinformatics. Newly beta-release Service-Finder [4] claims to have 25 thousand registered web services, however most queries produced handing error while authors were testing its capabilities. In both first two portals, results takes remarkable time to load. Moreover, many of the returned results are not related to the user need. For example, if a user is looking for a free weather forecasting web service, Seekda will retrieve web services with term "free" in the description without confirming that this web service is actually free to use or not. Moreover, the user need to read descriptions of retrieved web services in order to confirm if it is free or commercial. Authors are trying with this work to provide discovery processes with a classification solution to enhance discovery speed, and to help retrieve only classes of web services that have been selected by user during the search.

2. Related Work

Classification of web services is the act of grouping similar web services into groups. The similarity among a group of web services depends on different criteria, which leads to different classification methods. Classification enhances the speed of web service discovery process. Moreover, classification of web services increases the accuracy of discovering the right service for the specified need. Web services can be classified in different criteria. The following are the criteria being used in classification of web services in recent publications.

2.1. Behaviour

Classifying web services by the functionalities it provide. For example, informative web services is a category were web services provide information without required input from the consumer. For example, weather forecasting, money exchange rates, and global time services. Hongbing in [5] proposes an automated classification algorithm which depends on functional features such as inputs and outputs extracted from web services description files. Hongbing classification criteria depends on a standard taxonomy by UNSPSC.

2.2. Ontology and semantics

Crasso [6] proposed ontological based classification of web services inspired by vector space model. Each word from a web service description file has a weighting scheme based on TF-ITF scheme. His work shows flexibility in managing web services description, even though sources of words representing web services was not precisely determined. One defect of ontology based classification is the fact that different ontologies may cause different classification for the same web service, which may lead to ambiguity in web services discovery. Kehagias [7] proposed a semantic based classification of WSDL files through 3 layers of categorization. His work deployed WordNet [8] as ontology references and

PorterStemmer [9] for unique keywords extraction from WSDL description files.

2.3. Context Domain

Context domain classification has shown good results in clustering and grouping of web services. This is because consumers normally look for a web service by searching its domain. Abujarour et al [10] proposed an automatic classification algorithm which retrieve crawled web service description files from the web, stem their description and tags, then hash features of each web service using SimHash [11] function and classify web services depending on a domain classification extracted from programmableweb.com web page.

2.4. Quality of service (QoS)

Lee et al. proposed at [12] a web service quality management system WSQMS, which they have integrated with UDDI registry *tModel* component in order to link each web service with its quality of service parameters, and then used deployed these parameters in classifying web services. The problem of qos categorization is the lack of semantics, which is highly needed by discovery algorithm. However, QoS categorization is very helpful for web services clustering and filtering, which highly helps end user on making decision of what web service to choose among a group of similar functionality web services.

3. Proposed Classification Algorithm

Web services will be classified and stored into classes according to a non-functional criteria. These classes belongs to different criterions, which are predefined and provided by the web service registry. Fig 1 illustrates two non-functional criteria used in our proposed model. Classification attributes values are provided by web service provider during registration of the web service to the registry. However, for the experiment authors are building, and the examples illustrated later, classification of web services are generated randomly for each web service.

3.1. Definitions

This section will contain definitions of classification terms used in the classification criteria in this work:

- Free Unlimited: a web service is totally free and can be used for unlimited times.
- Free Limited: the web service is free for a limited time or number of usage times (trial).
- Subscription: the web service can be used only if the user is a subscribed customer, which means he pays some fee for some period of time to use the service, such as accessing a commercial magazine.
- Pay-per-use: the web service performs its functionality every time the customer pays for it, such as ordering flowers or buying software online.
- Collective: the web service has no output. Means that it does not respond to the end user with any information rather than a notification that the invocation was successful. An example of this type of web services is reporting web services.
- Notifying: in this case, the web service does not require any input from the end user in order to function. It only publishes information, such as weather forecasting and money exchange rates.
- Interactive*:* the web service receives input and produces output while interacting with the end user. An example of this category of web services is a credit card verification service, or a ticket booking service.

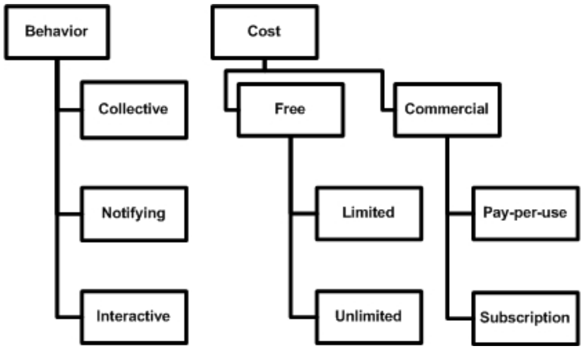


Fig. 1: non-functional classification criteria levels.

Our proposed classification model is meant to be supplied by web service provider by selecting the classification criteria that suits their web service while they are registering it at the web service registry. An algorithm will translate the web service provider selections into an id. This id will represent the non-functional classes of the web service. The algorithm gives each classification level two digits to be represented. Since the proposed classification has three levels, classification id will consist of 6 digits. Fig. 2 illustrates an example of a classification id generation.

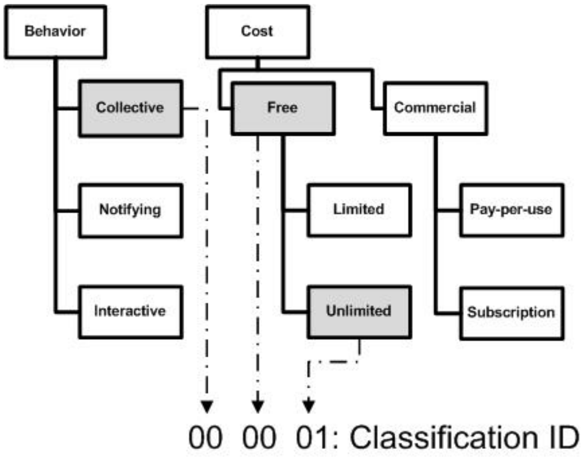
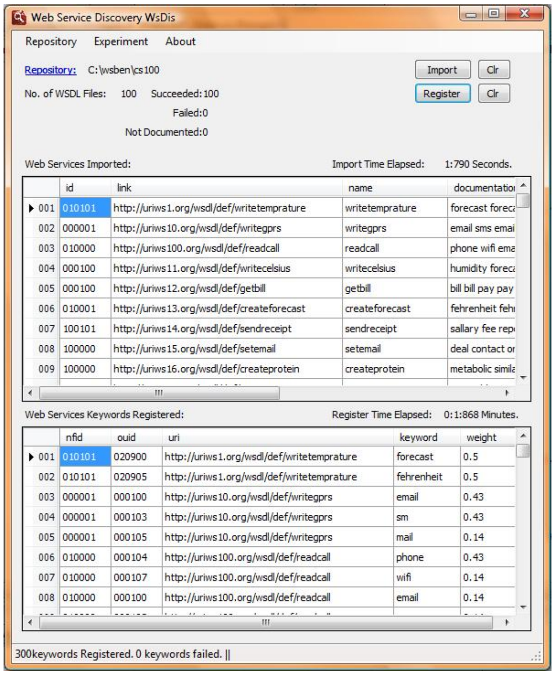


Fig. 2: classification ID generation example.

Considering three levels of classification, the highest level is behaviour-based, the second and third levels cost-based. Fig. 2 shows an example of generating a classification ID for a totally free and collective web service.

4. Experiment

In order to proof efficiency of proposed algorithm empirically, a prototype named WSDis was developed, which implements classification and discovery algorithms with QoS filtering model. Discovery algorithm and QoS filtering model are not reported in this paper.



Repository: C:\wsben\cs100

No. of WSDL Files: 100 Succeeded:100

Failed:0

Not Documented:0

Web Services Imported: Import Time Elapsed: 1:790 Seconds.

	id	link	name	documentatio
001	010101	http://uriws1.org/wsdl/def/writetemprature	writetemprature	forecast foreca
002	000001	http://uriws10.org/wsdl/def/writegprs	writegprs	email sms emai
003	010000	http://uriws100.org/wsdl/def/readcall	readcall	phone wifi ema
004	000100	http://uriws11.org/wsdl/def/writecelsius	writecelsius	humidity foreca
005	000100	http://uriws12.org/wsdl/def/getbill	getbill	bill bill pay pay
006	010001	http://uriws13.org/wsdl/def/createforecast	createforecast	fehrenheit fehr
007	100101	http://uriws14.org/wsdl/def/sendreceipt	sendreceipt	sallary fee rep
008	100000	http://uriws15.org/wsdl/def/setemail	setemail	deal contact or
009	100000	http://uriws16.org/wsdl/def/createprotein	createprotein	metabolic simila

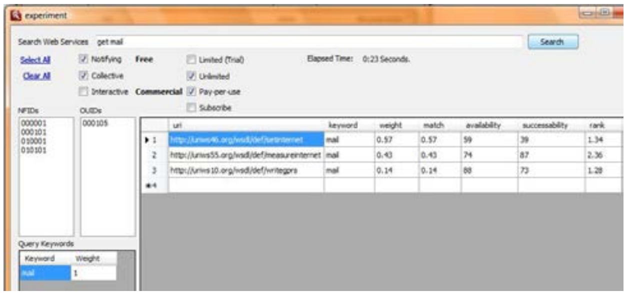
Web Services Keywords Registered: Register Time Elapsed: 0:1:868 Minutes.

	nfid	ouid	uri	keyword	weight
001	010101	020900	http://uriws1.org/wsdl/def/writetemprature	forecast	0.5
002	010101	020905	http://uriws1.org/wsdl/def/writetemprature	fehrenheit	0.5
003	000001	000100	http://uriws10.org/wsdl/def/writegprs	email	0.43
004	000001	000103	http://uriws10.org/wsdl/def/writegprs	sm	0.43
005	000001	000105	http://uriws10.org/wsdl/def/writegprs	mail	0.14
006	010000	000104	http://uriws100.org/wsdl/def/readcall	phone	0.43
007	010000	000107	http://uriws100.org/wsdl/def/readcall	wifi	0.14
008	010000	000100	http://uriws100.org/wsdl/def/readcall	email	0.14

300keywords Registered. 0 keywords failed. ||

Fig. 3 WSDis main form "Web Services Registration"

WSDis is developed in C# and it has two main stages, (1)Registration: were web service description files are categorised and indexed, the second stage is (2)Discovery: where a GUI is provided for searching the web services registered in the earlier stage.



Search Web Services get mail

☑ Notifying Free ☐ Limited (Trial) Elapsed Time: 0:23 Seconds.

☑ Collective ☑ Unlimited

☐ Interactive Commercial ☑ Pay-per-use

☐ Subscribe

NFIDs: 000001 000101 010001 010101

OUIDs: 000105

	uri	keyword	weight	match	availability	successability	rank
1	http://uriws46.org/wsdl/def/setinternet	mail	0.57	0.57	59	39	1.34
2	http://uriws55.org/wsdl/def/measureinternet	mail	0.43	0.43	74	87	2.36
3	http://uriws10.org/wsdl/def/writegprs	mail	0.14	0.14	88	73	1.28
4							

Query Keywords

Keyword	Weight
mail	1

Fig. 4 WSDis experiment form "Web Service Discovery"

Since experiment targets enhancement of speed, web services are distributed randomly under the classification criteria during registration stage. classification criteria was described earlier in this paper. Fig. 3 shows the main form of WSDis where web services description documents are imported and registered , and Fig. 4 shows search form where discovery stage takes place assisted with selection process, where classification algorithm functions.

4.1. Web Services Distribution and Coverage

This expression "web services distribution" implies the arrangement of web services in categories under the specified classification criterion.

The expression "web services coverage" implies the web services which are included in the discovery process which are included in the categories of classification criteria.

It is important to show the distribution of web services in order to relate the amount of time saving with the percentage of web services covered by selection process. Table 1 lists the classification distribution for the five datasets used in our experiment.

Table 1: Distribution of web services in classification criteria

Dataset Size:	100	500	1000	5000	10000
Collective	23%	31%	33%	34%	33%
Notifying	42%	35%	32%	31%	33%
Interactive	35%	32%	33%	33%	33%
Free Limited	35%	26%	23%	25%	24%
Free Unlimited	31%	25%	26%	24%	25%
Pay-per-Use	11%	24%	24%	25%	24%
Subscribe	23%	23%	25%	24%	25%

Size of dataset represents the number of WSDL files used. Classification criteria are Behaviour-based (Collective, Notifying, and Interactive) and Cost-based (Free limited, Free unlimited, Pay-per-use, and Subscribe). Numbers in Table 1 shows distribution in each criterion separated from the other.

4.2. Implementation

For each dataset, web services are imported and registered in the WSDis prototype, then a query was conducted 4 times:

- With one category selected.
- With two categories selected.
- With six categories selected.
- With all categories selected: means that the selection algorithm is not used and all web services are covered.

The same query were used in all runs over all datasets in order to make the comparison valid. For each run, the prototype records the time elapsed, which we compare each time with the 4th run result (all categories selected) in order to calculate the percentage of time saving. The following section discusses the results and compares experimental results with mathematical calculation.

5. Results and Discussion

Since our proposed classification is generic, authors find it necessary to prove its efficiency and usefulness mathematically. Then followed by an experiment conducted on an ontology-based discovery algorithm proposed by authors of this paper.

4.1. Mathematical Proof

In this section, authors will provide a mathematical proof for what it was claimed earlier, that the non-functional classification of web services will enhance discovery speed.

Hypothesis: non-functional classification of web services enhances discovery speed.

Proof: during search for a match to a value of a non-functional id of a web service record, if a match took place, a further processes should take place, otherwise, the matching process should continue and check the next record.

Let's say that the time required for checking a match/non-match status for a non-functional id is t_{nf}, and the time required for the discovery matching process to check a record t_{disc}.

This means that for each record, the time spent is either t_{nf} or $t_{nf} + t_{disc}$. Let's say that number of matched records is x and non matched records is y. Total time used to run through all records and find matching records can be calculated in the following formula:

$$\text{Time} = x * (t_{nf} + T_{disc}) + y * (t_{nf}) \qquad (1)$$

Time spent for discovery matching process is definitely bigger than time used for non-functional matching. This is due to the reasoning nature of matching and the involvement of several components in discovery algorithms, especially ontology reasoning or novel proposed additional web service descriptions like semantic descriptions. However, we consider the simplest discovery matching process may require time at least equal to the non-functional matching process. Having this consideration in mind, we can simplify the previous formula making $t_{nf} \approx t_{disc} = t$.

$$\text{Time} = x * (2t) + y * (t) \qquad (2)$$

Now, let us calculate the time needed to check matching of both non-functional and discovery matching processes. For each record, there will be t_{nf} and t_{disc}. Which means:

$$\text{Time} = (x + y) * (2t) \qquad (3)$$

Time saving by deducting Eq. (2) from Eq. (3) can be calculated as in Eq. (4) as follows:

$$\text{Time saving} = (2xt + 2yt) - (2xt + yt) = yt \qquad (4)$$

Since y > 0, and t > 0, thus, time saving > 0. In case y = 0, it means that user decided not to user non-functional classification during discovery, which yields the system will go through all records matching non-functional and ontology. It is obvious that there will be no time saving. Percentage of time saving out of total time can be calculated by Eq. (5). TSP is the abbreviation for Time Saving Percentage as.

$$\text{TSP} = \frac{\text{time saving}}{\text{whole time}} = \frac{Yt}{2xt + yt} = \frac{Y}{2x + y} \qquad (5)$$

Fig. 5 shows the relation between web service classification coverage and time saving. Classification coverage means the number of web services included in the discovery process using selected classification criteria by the user.

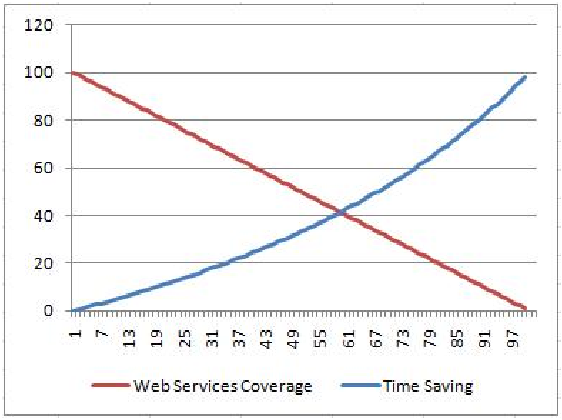


Fig. 5 Relationship of non-functional classification coverage and time saving.

Coverage is inversely proportional with time saving as Fig. 5 illustrates. However, experimental results depends on the distribution of web services among the categories, and real relation should vary slightly from Fig. 5. as Table 1 shows. In the next section, experimental results will be exposed.

4.2. Experimental Results

Five datasets of web services were generated using a customized version of WSBen [13] framework. Authors have made the customization to WSBen in order to generate web services with meaningful name and documentation. dataset 1 contains 100 web service WSDL description files, dataset 2 contains 500, dataset 3 contains 1000, dataset 4 contains 5000, and dataset 5 contains 10000 WSDL files. Each dataset was deployed four times, each time with different classification coverage. First time with 1 category covered, second time with 2, third time with 6, and the last time with all categories included (this is equal to not using classification at all for discovery). Fig. 6 illustrates time saving in the 5 different datasets mentioned earlier.

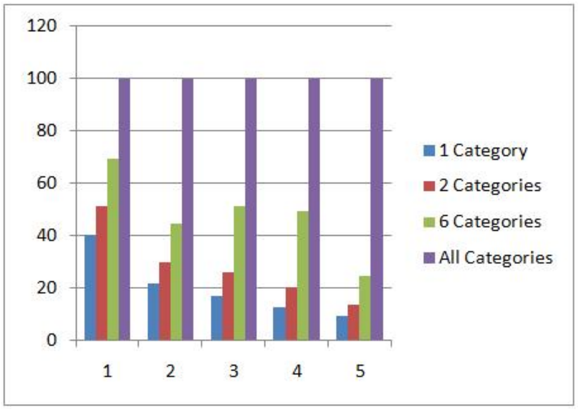


Fig. 6: time saving for different non-functional classification coverage.

Results have shown remarkable saving of time - 50% to 90% - when using one or two categories, especially when the number of web services is big, such as in dataset 5 of 10000 web service description files. this is due bypassing the reasoning logic of matching, which consumes nontrivial time in most cases of discovery algorithms.

6. Conclusion

Number of web services that being publically available is increasing tremendously recently, and the discovery process speed issue is coming to the surface. Classification of web services adds remarkable benefits to web service discovery processes. It helps categorize web services and contributes to the discovery process speed enhancement. Our proposed non-functional classification depends on three layers of classification criteria. Mathematical and experimental proofs has been conducted, and they have validated that our proposed classification is useful and efficient.

References

1. Halima, R.B., K. Drira, and M. Jmaiel. *A QoS-driven reconfiguration management system extending Web services with self-healing properties*. in *Proceedings of the Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises, WET ICE*. 2007. Paris.
2. Kritikos, K. and D. Plexousakis. *Semantic QoS-based web service discovery algorithms*. in *Proceedings of the 5th IEEE European Conference on Web Services, ECOWS 07*. 2007. Halle.
3. Bhagat, J., et al., *BioCatalogue: a universal catalogue of web services for the life sciences*. Nucleic Acids Research, 2010. **38**(suppl 2): p. W689-W694.
4. *Service Finder portal*. [cited 2011 10 June]; Available from: www.service-finder.eu.
5. Hongbing, W., et al. *Web Service Classification Using Support Vector Machine*. in *Tools with Artificial Intelligence (ICTAI), 2010 22nd IEEE International Conference on*. 2010.
6. Crasso, M., A. Zunino, and M. Campo, *Combining Document Classification and Ontology Alignment for Semantically Enriching Web Services*. New Generation Computing, 2010. **28**(4): p. 371-403.
7. Kehagias, D.D., et al., *A WSDL Structure Based Approach for Semantic Categorization of Web Service Elements*, in *Artificial Intelligence: Theories, Models and Applications, Proceedings*, S. Konstantopoulos, et al., Editors. 2010, Springer-Verlag Berlin: Berlin. p. 333-338.
8. Miller, G.A., *WordNet: a lexical database for English*. Commun. ACM, 1995. **38**(11): p. 39-41.
9. Porter, M. *Porter Stemming Algorithm*. 2006 2006 [cited 2010; Available from: http://tartarus.org/~martin/PorterStemmer/.

10. AbuJarour, M., F. Naumann, and M. Craculeac, *Collecting, annotating, and classifying public web services*, in *Proceedings of the 2010 international conference on On the move to meaningful internet systems - Volume Part I*. 2010, Springer-Verlag: Hersonissos, Crete, Greece. p. 256-272.
11. Charikar, M.S., *Similarity estimation techniques from rounding algorithms*, in *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*. 2002, ACM: Montreal, Quebec, Canada. p. 380-388.
12. Lee, Y., *Quality-context based SOA registry classification for quality of services*, in *Proceedings of the 11th international conference on Advanced Communication Technology - Volume 3*. 2009, IEEE Press: Gangwon-Do, South Korea. p. 2251-2255.
13. Oh, S.C. and D. Lee, *WSBen: A web services discovery and composition benchmark toolkit*. International Journal of Web Services Research, 2009. **6**(1): p. 1-19.

Mamoun Mohamad Jamous received his Bachelor degree in computer science from Sudan University for Science and Technology at 2005. Msc in Computer Science at Universiti Teknologi Malaysia at 2008. He is currently a PhD student at Universiti Teknologi Malaysia. His research interest is Web Services Discovery, Software Architecture, Software Requirements, and Computer Architecture.

Safaai Bin Deris is a full professor of Computer Science at Universiti Teknologi Malaysia. His research interest is Software Engineering, Service-Oriented Architecture, Bioinformatics, and Scheduling, and Artificial Intelligence computing.

Mingling Multipath Routing With Quality Of Service

Dr. Shuchita Upadhyaya[1] and Gaytri Devi[2]

**1.Department of Computer Science and Applications
Kurukshetra University
Kurukshetra**

**2.GVM Institute of Technology and Management
DCRUST , Murthal Road
Sonipat**

Abstract

The QoS issue in the Internet have become essential for the successful transmission of multimedia applications . The basic problem of QoS routing is to find a path satisfying multiple constraints. It is concerned with identifying the path that will consider multiple parameters like bandwidth, delay, cost, hopcount etc. instead of one .To provide user- or application-level Quality of Service (QoS) guarantee Multipath routing strategy can be used for the transmission of QoS sensitive traffic over the network. Multipath routing means using multiple paths instead of using single path to forward the traffic.If multiple paths are being used for the transmission of the traffic then the traffic will be redirected to the back up path if active path fails. In this way robustness can be achieved. On the other hand load balancing for communication network to avoid network congestion & optimize network throughput also requires multi paths to distribute flows . Robustness & load balancing are aspects of QoS routing . So multipath can be proved very valuable for Quality of service. This paper investigates the approaches of mingling Multipath & Quality of service. The approaches considered are based on Dijkstra algorithm, Bellman ford algorithm, Resource reservation & MPLS.

Keywords: *Multipath routing, Multiple paths, Single shortest path, Quality of Service.*

1. Introduction

In today's era of Internet ,the demand of real time multimedia applications have been increased .To fulfill this requirement, Quality of Service(QoS) factors have become necessary to be present in the network. For example, transmission of video over a computer network should be without undesirable delays and jitter and a medical image or a robot control packet may be required to be transmitted over a network with the minimum end-to-end delay. The present Internet routing mechanisms (based on the best-effort paradigm) are unlikely to provide such end to- end performance guarantees required in these applications. Here is a need of the mechanism which will consider these factors(delay ,jitter, bandwidth etc.) for the transmission. One of the components of that mechanism is QoS routing. Multipath approach can be merged into QoS Routing to catch its maximum advantage as in some situation a single path is not able to fulfill all the QoS requirements.

The benefits of provisioning multiple QoS paths are reliable QoS support and uniformly balanced network load. In addition to ground network, it is also important to provide QoS support in the presence of network failures. The approach is to take advantage of multiple alternate paths in the face of network failures. The scheme shows major improvement of fault tolerance .

In this paper, we have discussed four approaches of multipath algorithms based on QoS criteria. The approaches are based on Dijkstra , Bellman Ford ,Resource reservation & MPLS. The layout of the paper is as follows: Section 2 describes Multipath routing. Section 3 describes Quality of service. Section 4 discusses approaches of combining QoS with Multipath. Section 5 concludes the paper.

2. Multipath Routing

Unlike traditional routing schemes that route all traffic along a single path ,multipath routing strategy uses multiple paths. Single path routing may lead to unbalanced traffic distribution and congestion. It can not achieve the proper utilization of resources. In contrast to single path approach, multipath routing can better utilize network bandwidth and balance network traffic. There are two strategies for allocating traffic over available path .First is to distribute traffic among multiple paths instead of routing all the traffic along a single path. Second is to forward traffic using only the path with the best metric and keep other discovered paths as backups that can be used in case of congestion or blocking. Thus multipath routing is an alternative to single shortest path routing to distribute load and lighten congestion in the network .

2.1Benefits of Multi path Routing

Multipath routing would offer many benefits as following-

Fault Tolerance : When multiple path are available ,traffic can move to an alternate path on the occurrence of congestion. This will lead to less delay and packet loss.

Increased Bandwidth: If multiple path exists ,an application can access more bandwidth by using multiple path simultaneously. As multipath routing has the potential to aggregate bandwidth allowing a network to support data transmission rates higher than what is possible with any one path.

Improved Reliability: If multiple path exists ,traffic can switch quickly to an alternate path when a link or router fails.

Load Balancing: By using multiple path simultaneously ,network resources can be more used by distribution of traffic among several paths . This is reverse to single shortest path routing scheme where one path is completely busy and others are under loaded . So with multipath routing load balancing can be achieved.

3 . Quality Of Service

The fundamental problem of routing in a network that provides QoS guarantee is to find a path between specified source and destination node pair that simultaneously satisfies multiple QoS Constraints such as cost, delay & reliability.

Quality of Service(QoS) puts some restrictions in the form of certain constraints on the path. These constraints may be desired bandwidth, delay, variation in delay experienced by receiver(jitter),packet loss that can be tolerated, no of hops, cost of links etc.

QoS Constraints are represented in the form of metrics. One metric for each constraint is to be specified like bandwidth metric, jitter(variation in delay) metric, delay metric, no of hops metric, packet loss ratio etc. for one node to all other nodes in the network. Metric for a complete path with respect to each parameter is determined by the composition rules of metrics.

By combining a set of QoS metrics in a single metric, it is possible to use existing polynomial-time path computation algorithms, such as Bellman–Ford or Dijkstra. The three basic rules are-

Additive Metric: The value of that constraint for a path is the addition of all links constituting path. For Example- delay, hop count, cost, jitter.
It can be represented as

$$D(pi\ )=\sum_{e \varepsilon\ pi}(d(e))$$

It means delay of path is sum of all its edges.

Multiplicative Metric: Using this metric, The value for the complete path is multiplication of all its edges .
Examples are – reliability(1-lossratio) and error free Transmission (probability)
It can be represented as

$$R(pi\ )=\prod_{e \varepsilon\ pi}(r(e))$$

The reliability of the path is multiplication of all its edges. Multiplicative metric can be converted into additive by taking logarithm.

Concave Metric: In this metric, either we can take min value or max value among all the edges for a path. For Example- Bandwidth
It can be represented as-

$$B(p)=min/max\ (b(e))$$

For a complete path, the constraints may be required either as a constrained form or in a optimization form. In constrained form, some condition is put on constraint value e.g. Choose that path only which has delay less than or equal to 60 ms. The path obeying the condition is called feasible. On the other hand optimization refers to path having minimum or maximum value for a constraint e.g. Choose the path that has minimum delay among all the paths. This path is called optimal path .The further QoS issues have been discussed in[4].

4. Multipath Approach And Quality Of Service

The two main concerns to implement multipath routing scheme are the calculation of multiple paths and traffic distribution among multiple paths. To determine the multiple paths various k-shortest path algorithms have been used. The k-shortest means determining not only the shortest ,but also the second, the third …………the kth shortest path (for given integer k>1).Regarding this, two different types of problems are usually considered : the unconstrained and the constrained k-shortest path problem. While in the former no restriction is considered in the definition of path. In the constrained k shortest path problem all the paths have to satisfy some condition . e.g. to be loop less and to be disjoint. A path from s to t is a loop less path, if all its nodes are different . A path is disjoint if it is link disjoint or node disjoint. Paths between a given pair of source and destination nodes in a network are called link disjoint if they have no common (i.e. overlapping) links and A pair of paths is considered node-disjoint if, besides the source and destination nodes, they have no common nodes. Node disjoint paths provide more reliability then the link disjoint paths.

In general a link-disjoint paths algorithm can be extended to a node-disjoint algorithm with the concept of node splitting, i.e. replacing one node with two nodes that are linked together via a link with zero-valued weights[15].

To find the k-shortest path, shortest path algorithms Dijkstra & Bellman-ford-Moore algorithms can be used in the generalization form [5]. The K-shortest paths are limited to defining alternate paths without consideration of QoS constraint.

So here is a need of enhancing the above mentioned algorithms to consider QoS parameters.

The main Multi path QoS approaches to provide multiple QoS paths are as follows-

4.1 Using BellMan ford algorithm-

In considering with multiple constraints, it has been noticed that the BF algorithm can potentially solve a two metric routing problem [1].It is a property of the BF algorithm that, at its h-th iteration, it identifies the optimal path between the source and each destination, among paths of at most h hops .It searches for a minimum path cost (or maximal bandwidth) in ascending order of no of hops .The cost of a path is a function of its available bandwidth i.e. the smallest available bandwidth on all links of the path, and finding a minimum cost path amounts to finding a maximum bandwidth path. However, because the BF algorithm progresses by increasing hop count, it inherently provides for the hop count of a path as a second optimization criteria.

So the result of the algorithm comes with the path that is the one with maximal available bandwidth among all the feasible paths with the minimum no of hops.

.

Thus Bellman Ford is very powerful in solving most multiple constrained routing problem if the minimum hop is the main objective function. BF algorithm is also capable of solving delay ,delay jitter, loss & bandwidth constrained routing problem[2].This Bellman ford algorithm is single QoS path computation algorithm. This Single path computation algorithm has been extended to perform multiple QoS computations [3].Each of the QoS metric is manipulated in the same way as in single path algorithm by increasing hop count. The algorithm has been designed in order to improve fault tolerance & load balancing . The multiple path computation algorithm searches for maximally disjoint paths(minimally overlapped paths) so that impact of link failure is reduced & links are more evenly utilized by spreading the network load over multiple paths .In order to search for multiple alternate paths to provide fault tolerance and load balancing yet satisfying QoS constraints ,it has defined alternate paths with the following conditions –
-Satisfying given QoS constraints
-Maximally disjoint from already computed paths
-Minimizing hopcount.

The algorithm searches for multiple maximally disjoint paths (i.e with the least overlap with each other) such that the failure of a link in any of the paths will still leave (with high probability) one or more of the other paths operational

Based on this algorithm [13] has presented the multiple QoS path algorithm with some enhancements for maximally disjoint paths i.e PDMA (partially disjoint multiple QoS path algorithm with multiple iteration)and also presented fully disjoint multiple QoS path algorithm (FDMA).

4.2 Using Dijkstra Algorithm

Dijkstra algorithm can generate a minimum hop path that can accommodate the required bandwidth and also has maximum bandwidth. But it requires some modification in order to be able to calculate the minimum hop path computation as it does not search for a minimum path cost in ascending order of no of hops as Bellman-Ford does. The modification required for supporting them is straightforward. Firstly on a graph from which all edges, whose available bandwidth is less than that requested by the flow triggering the computation, have been removed. This can be performed either through a pre-processing step, or while running the algorithm by checking the available bandwidth value for any edge that is being considered. Another modification to a standard Dijkstra based minimum hop count path computation, is that the list of equal cost next (previous) hops which is maintained as the algorithm proceeds, needs to be sorted according to available bandwidth. This is to allow selection of the minimum hop path with maximum available bandwidth. Alternatively, the algorithm could also be modified to, only keep among equal hop count paths the one with maximum available bandwidth. This would essentially amount to considering a cost that is function of both hop count and available bandwidth[1].

Bellman ford algorithm used in computing k shortest paths ignores equal hop multiple count but equal hop multiple paths impairs the performance of routing protocol if the link state information is inaccurate

[6][9] have presented a bandwidth constrained multipath routing algorithm based on Dijkstra . Besides specifying the optimality criteria that define the best paths , it has also described the construction & selection strategies for multiple paths. The path construction is performed each time link state information becomes available , while the path selection is performed for each connection request .

This algorithm is based on k best one to one non loop less paths. It has modified the algorithm to find k one to all loop less paths instead of one to one non loop less paths. The algorithm generates k paths that are either shortest in terms of number of hops(hop based) or widest in terms of bottleneck bandwidth(bandwidth based).It has considered 2 metrics bandwidth & hop count. It has developed 2 categories of k-shortest algorithms-hop based & bandwidth based & five selection algorithms. It has concluded that hop based algorithm outperforms the bandwidth based one. It has also presented five path selection schemes as Best-K-Widest(BKW),Random-K-Widest(RKW),Shortest-K-widest(SKW),Best-K-Shortest(BKS),Widest-K-Shortest(WKS).

4.3 Using Resource Reservation

In order to use the network resources efficiently, bandwidth reservations are made to ensure high probability of data arrival to its destinations.

Bandwidth reservation means If certain amount of bandwidth is reserved for a program and then when the program accesses the network, up to that much bandwidth is guaranteed to be available to the program.

By reserving bandwidth, it is possible to provide reasonable levels of QoS. The idea is to identify traffic flows, which are streams of packets (voice, video, file transfers, and so on) going to the same destination IP address and port number. Reservations are negotiated with each network device along a route to a destination. If each device has resources to support the flow, a reserved path is set up.

[16] has presented a QoS routing algorithm in which the QoS constraint is specified by bandwidth guarantee. Therefore, the goal of QoS routing algorithm is to find the best path with sufficient bandwidth. Here the best path as the path with least cost. To find the least cost path, the algorithm can run either Dijkstra or Bellman-Ford shortest path algorithm.

[11] has presented a family of algorithms that route and reserve resources along parallel sub routes i.e. fast algorithm ,slow algorithm super fast algorithms.

[12] has presented a framework in which bandwidth can be reserved on the communications links, and, once reserved, is guaranteed for the required time period. The algorithms finds multi paths, consisting of possibly overlapping paths, based on the available bandwidths on various links of the network. It considers delay as a second consideration criteria. It has presented the algorithms based on two problems. The first problem requires that a message of finite length be transmitted from s to d within r units of time. The second problem requires that a sequential message of r units be transmitted at a rate of n such that maximum time difference between two units received out of order is no more than q.

4.4 Algorithms for MPLS

Multi-Protocol Label Switching (MPLS) networks do routing based on connections . QoS routing and traffic engineering goals are normally achieved by finding optimal or near-optimal explicit routing MPLS algorithms .The MPLS can provide fast packet forwarding .It is a new internet technology that is rapidly emerging as a core technology for next generation networks . MPLS uses a technique known as label switching to forward the data through the network. Multi-protocol label switching (MPLS) has many attractive features. First of all, the MPLS can create easily an explicit-route label switched path as needed and it can easily map traffic trunks that consist of traffic flows with similar characteristics of traffic requirements.

[10]Proposed two multipath constrained based routing algorithms using MPLS. The algorithms proposed in this paper find multiple LSPs between ingress and an egress LSR satisfying a given bandwidth constraint. When there is

no single path through the network satisfying a whole bandwidth constraint, the suggested algorithms divide the bandwidth constraint into two or more sub-constraints and find a constrained path for each sub-constraint. First, the algorithms calculate the least-cost path between source and destination, and allocate the path bandwidth (minimum bandwidth of all links along the path) to that path. Next, it calculate the next shortest path through the network after removing links having no available bandwidth and allocate the path bandwidth to that path. The process is continued until it can allocate the whole bandwidth constraint to the successive shortest paths. If a single path satisfies the whole bandwidth constraint, there is no need to balance the traffic load. However, if there are multiple paths the traffic load should be partitioned optimally for mapping to multiple paths.

[14] proposed a new constraint based routing algorithm for MPLS networks. It has used bandwidth & delay constraint. In this algorithm the best path is selected based on Best –fit strategy. In step1, the algorithm eliminates all the links with bandwidth value less than the bandwidth constraint. In step2 ,it finds the path with delay value less than (or equal to) delay constraint value. In step3 , if the path with minimum hop count is used, the minimum of network resources are consumed. In step 4, the load is distributed & balanced. The best fit load balancing strategy selects path from all feasible paths which has the nearest bandwidth value to the bandwidth constraint value.

[17] proposed a localized proportional routing approach where each source node collects information about the traffic originating from itself and computes proportions based solely on this local information. Global schemes have to gather system wide traffic metrics and thus slower to react to changes. Localized schemes , on the other hand use only local information and thus can adapt to change faster. It has assumed that flow from source to destination arrive randomly with a Poisson distribution and their holding times are exponentially distributed. The algorithm has used the strategy equalization of blocking probabilities (ebp) of candidates paths. This ebp strategy requires only path level information.: the amount of offered load and the corresponding blocking probability. The objective of ebp strategy is to find a set of proportions such that flow blocking probabilities on all the candidate paths are equalized. The strategy can be implemented using the following procedure to compute new proportions in each iteration .First, the current average blocking probability is computed. Then the proportion of load onto a path is decreased if its current blocking probabilities higher than the average and increased if lower than the average. When we talk about path selection, the algorithm selects the candidate paths of a pair that do not share bottle neck links .In this way of path selection, the blocking probability can be reduced. To judge the goodness of paths, It has introduced the notion of width for a set of paths, which is defined as the maximum flow carriable by paths in the set. The amount of flow carriable by a link is given by its average available bandwidth. So the width of a set of paths can be computed given the average available bandwidth information about each link in the network. Based on the notion of width of a path set, the algorithm proposed a path selection procedure that adds a new candidate path only if its inclusion increases the width. It deletes an existing candidate path if its exclusion does not decrease the total width. So this proportional routing scheme yields higher throughput with lower overhead.

5. Conclusion

The multiple paths are utilized in parallel making the entire network system less prone to network failure. Multipath approach can be used as an architecture for implementing quality of service by aggregation of flows. In connection oriented networks with QoS guarantees ,they reduce blocking probabilities. By provisioning multiple QoS paths, the network system can provide backup paths when one or more paths are detected as corrupted. Besides, the spreading of network traffic over the provisioned multiple QoS paths favors even network resource utilization

There are various ways to combine QoS and Multi path approach. This combination should be further enhanced for the current need of the Internet applications .

References

[1] R. Guerin, A. Orda, D. Williams, “QoS routing mechanisms and OSPFextensions”, Proceedings of Global Internet (Globecom), Phoenix,Arizona November (1997).

[2] D. Cavendish, M. Gerla, “Internet QoS routing using the Bellman–Ford algorithm”, IFIP Conference on High Performance Networking(1998).

[3] S.S. Lee, M. Gerla, “Fault tolerance and load balancing in QoS provisioning with multiple MPLS paths”, Lecture Notes n Computer Science, vol. 2092, International Workshop on Quality of Service(IWQoS), 2001, pp. 155.

[4] S.Upadhaya,G.Devi , “Characterization of QoS Based Routing Algorithms” , International Journal of Computer Science & Emerging Technologies 133 Volume 1, Issue 3, October 2010.

[5] E.Q.V. Martins, M.M.B. Pascoal and J.L.E. Santos, “The K Shortest Paths Problem,” Research Report, CISUC,June 1998.

[6] R. Guerin and A. Orda, “QoS-based Routing in Networks with Inaccurate Information: Theory and Algorithms.”

IEEE/ACM Pansaction on Networking, Vol. 7, No. 3,June 1999, pp. 350-364.

[7] H.Jiayue ,J.Rexford "Towards Internet –wide Multipath Routing" .

[9] Yanxia Jia, Ioanis Nikolaidis and P. Gburzynski, "Multiple Path QoS Routing",Proceedings of ICC'01 Finland, pages 2583-2587, June 2001.

10] H.Y.Cho, J.y Lee and B. C. Kim ,"Multi-path Constraint-based Routing Algorithms for MPLS Traffic Engineering,",March 2003 IEEE, pp. 1963-1967.

[11] I. Cidon, R. Rom, Y. Shavitt, "Multi-path routing combined with resource reservation", Kobe, Japan April (1997) 92–100.

[12] N.S.V. Rao, S.G. Batsell, "QoS routing via multiple paths using bandwidth reservation", San Francisco, California March/April(1998) .

[13] S.S. Lee, S. Das, H. Yu, K. Yamada, G. Pau, M. Gerla "Practical QoS network system with fault tolerance", Computer Communications 26 (2003) 1764–1774

[14] M.H. yaghmae ,A.A. Safaeei,,"Quality of service in MPLS networks using delay and bandwidth constraints".

[15] X. Masip-Bruina,, M. Yannuzzib, J. Domingo-Pascuala, A. Fonteb, M. Curadob,E. Monteirob, F. Kuipersc, P. Van Mieghemc, S. Avalloned, G. Ventred, P. Aranda-Gutie´rreze, M. Hollickf, R. Steinmetzf, L. Iannoneg, K. Salamatian, "Research challenges in QoS routing", Computer Communications 29 (2006) 563–581.

[16] Y.W. Chen, R.Hwang, Y-D. Lin, "Multipath QoS Routing with Bandwidth Guarantee", Jan-2001 IEEE pp.2199-2203.

[17] S.Nelakuditi and Z.L. Zhang ,"A Localized Adaptive Proportioning Approach to QoS Routing", IEEE Communications Magazine , June 2002 ,pp 66-71.

Dr. Shuchita Upadhyaya is an associate professor in the Department of Computer Science and applications in Kurukshetra University, India. She is Ph.D. in Computer Science and application. Her area Of specialization is Computer Network .She has published more than 45 research papers in various national and international Journals. She is having 23 years of teaching and research experience.

Gaytri Dhingra is an assistant professor in GVM Institute of Technology and Management , Sonipat, India .She has done MCA , M.Phil. Now she is pursuing Ph.D. in Computer Science and Applications from the department of Computer Science and applications, Kurukshetra University. She has published five research papers in various national & international journals. She is having more than 10 years of teaching experience.

Comparative Analysis for Discrete Sine Transform as a suitable method for noise estimation

Swati Dhamija[1] and Dr. Priyanka Jain[2]

[1] Quality Engineer, Aircom International, Gurgaon

[2] Assistant Professor, Delhi Technical University, New Delhi

Abstract

When the speech signal corrupted by noise is processed using different transforms like the Discrete Fourier Transform, Discrete Cosine Transform and the Discrete Sine Transform , a comparative analysis proves that the Discrete Sine Transform (DST) is most suitable for de-noising and therefore reconstruction of the original speech. The experimental results show that the discrete sine transform coefficients at the low frequency regions are predominantly speech, and at the high frequency regions are predominantly noise. According to this, a new noise estimation method based discrete sine transform is proposed in this paper. The usage of DST coefficients reveals that the energy distribution throughout the segment of speech is uniform.

Keywords: *Digital Speech Processing, Discrete Sine Transform, Discrete Cosine Transform, Noise Estimation.*

1. Introduction

Digital speech processing is an important part of the communication domain, in the course of speech technology. However, because of the presence of noise, the speech which the receiver gets is not the original speech. In order to obtain as pure as possible speech signal, we need to estimate the noise, but, because the noises are root in a large number of origins and are a mass of kinds, it is almost impossible to obtain the completely pure speech from the noised speech. The mainstream algorithms for speech enhancement nowadays are mainly based on Fourier Transform (DFT).

DST, DFT, DCT as well as K-L transform are reduced into an identical transformation family in comparison with the first order Markov process called the “sine family”. This means that the DST has the similar nature with these three kinds of transformations. DFT, DCT as well as K-L transformation are all applied to the speech enhancement and there is no correlated literature on sine transform.

The binding energy ability of DST is not as strong as DFT and DCT but that does not imply that the strongest is the best. The experiment shows that DST is better in the processing of white noise as compared to DFT and DCT.

2. Definition and Characteristics of DST

The Discrete Sine Transform (DST) is a Fourier-related transform similar to the discrete Fourier Transform (DFT), but using a purely real matrix. It is equivalent to the imaginary parts of a DFT of roughly twice the length, operating on real data with odd symmetry (since the Fourier transform of a real and odd function is imaginary and odd), where in some variants the input and/or output data are shifted by half a sample.

For an assigned sequence x(n), its DST transform and inverse transform are defined separately as:

$$X[k] = \sqrt{\frac{2}{N+1}} \sum_{n=1}^{N} x(n) \sin \frac{nk\pi}{N+1} \tag{1}$$

$$x(n) = \sqrt{\frac{2}{N+1}} \sum_{k=1}^{N} X[k] \sin \frac{nk\pi}{N+1} \tag{2}$$

Where n=1, 2, 3…N; k=1, 2, 3…N.

DST, which was originally, developed by Jain (type I) and Kekra and Solanka (type II), belong to the family of unitary transform. This unitary family includes the DST, discrete cosine transforms (DCT) and the discrete Fourier transforms (DFT).

Figure 1 presents the waves of a length of speech signal and the waves after three kinds of transforms, for convenience of comparison only the absolute parts are drawn for DFT. We can see from the figure that DCT is the best in binding energy. The transformed speech energy mainly concentrates on the low frequency area and the noise mainly concentrates on the high frequency area.

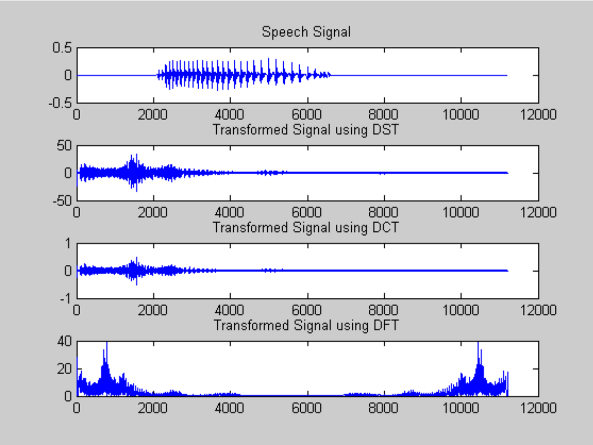


Figure 1. Speech signal transformations

3. Reconstruction of Original Speech

The inverse of the transformed signal results in the signal itself. However, the conversion of the signal to its transform and then inverse of the transformed signal to reconstruct the original speech waveform causes a shift in the coefficients thereby introducing noise in the original signal. The speech signal mainly concentrates on the low frequency part, and the noise energy mainly concentrates on the high frequency unit, so the estimation of noise is mainly done the high frequency area.

Figure 2 represents the waves of the reconstructed speech signal after performing the three inverse transformations.

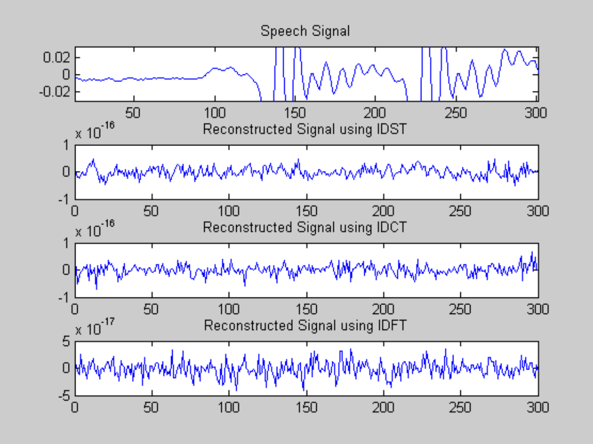


Figure 2. Reconstructed speech

3.1 Calculation of SNR

After the reconstruction of the signal using the three different transforms, we can calculate the Signal-to-Noise Ratios (SNR).The lower the signal to noise ratio is, the more accurate the noise estimate is.

We can compute the SNR as follows, where x(n) is the clean speech, $\tilde{x}(n)$ is the de-noised speech:

$$SNR = 10 * \log \frac{\sum_{n=0}^{N-1} x(n)^2}{\sum_{n=0}^{N-1}[x(n)-\tilde{x}(n)]^2} \qquad (3)$$

The computed values of SNR for the three transforms are as follows for a speech signal with length N, which is set to 11220.

Table 1: SNR Comparisons

Transformation Used	SNR value
DST	683.3081
DCT	703.4444
DFT	701.4479

The SNR values computed in Table1 imply that DST has the lowest value. This proves that the noisy components are maximum due to sine transformation, thereby making it suitable for noise estimation.

3.2 Noise Estimation using DST

For noise estimation, take the discrete sine transform on signal with length L, then take estimation in the last N/2 sine coefficients, that is the high frequency region.

$$\sigma_{v1}^2 = Var(ThelastN / 2ofDSTcoefficients) \qquad (4)$$

The discrete sine transform is an orthogonal transformation, so it cannot keep the energy of the signal to be invariable after transformation; therefore the variance we estimated from (4) is not the true noise variance. We add the white Gaussian noise to a speech with length L, which L is set to 11220, then take the DST of the signal and compute the estimated noise by the formula:

$$\sigma_v^2 = 2\sigma_{v1}^2 / L \qquad (5)$$

By estimating the noise we get the following data:

Table 2 Estimated Variance and Errors.

SNR	Actual Variance	Estimated Variance	Error
-15	9.0092e+004	16.0593	9.0076e+004
-10	2.7287e+004	4.8640	2.7282e+004
-5	9.1577e+003	1.6324	9.1561e+003
0	2.8220e+003	0.5030	2.8215e+003
5	879.5745	0.1568	879.4177
10	283.8823	0.0506	283.8317
15	88.1087	0.0157	88.0930

The set of experiments done in Table2 were then repeated for DCT and DFT. Figure3 shows the noise estimation using the three different transformations.

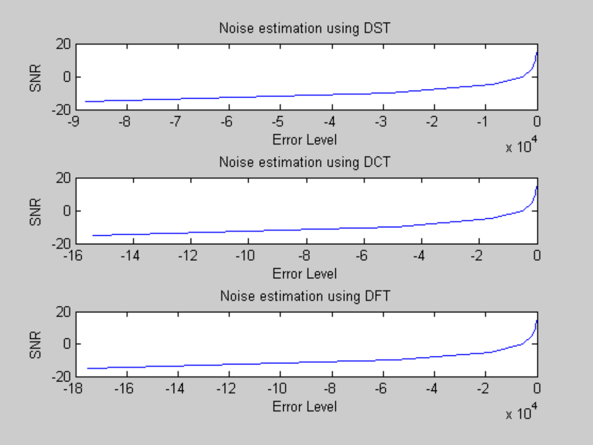


Figure 3. Noise estimation using various transformations

Figure4 is the error chart between estimated noise level and actual noise level under different signal to noise ratio, it can be seen from the chart that the lower the signal to noise ratio is, the more accurate the noise estimation is.

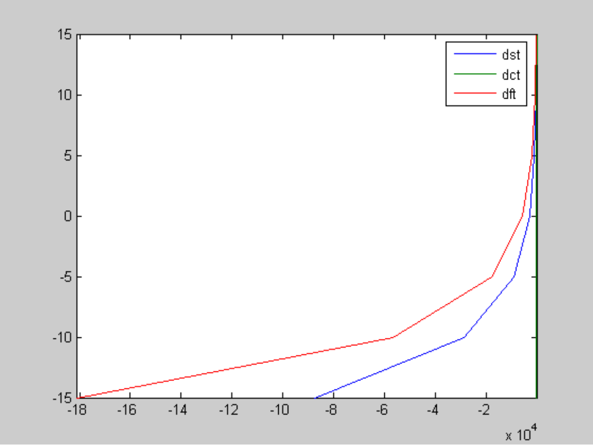


Figure 4. Error chart

4. Experiment

The proposed estimation method was tested on the speech data of the word “Hi”, spoken by a male. The segment length of the speech signal was about 10ms. The speech data used was sampled at 11220 Hz.
The speech signal was the processed using different transformations like the DST, DCT and DFT. Inverse transformations were then applied to the transformed signals to reconstruct the original speech. Signal to noise ratios were computed for all the three waveforms. White Gaussian noise of variable SNR values were then added to the original speech for noise estimation.

5. Conclusions

This paper adopts discrete sine transform as the method that generates better results for noise estimation as compared to DCT and DFT. The improvement is noticeable for white noise. In the future research, speech reconstruction using DST can be considered; the further improvement of the effect of de-noising may be seen.

References

[1] Jain AK. , “Fast Karhunen–Loeve transform for a class of stochastic processes”, IEEE Transactions on Communications1976; COM-24:1023–1029.

[2] Kekra HB, Solanka JK.,“Comparative performance of various trigonometric unitary transforms for transform image Coding”, International Journal of Electronics 1978; 44:305–315.

[3] Proakis JG, Manolakis DG, Digital Signal Processing (3rd edn). Prentice-Hall: Englewood Cliffs, NJ, U.S.A., 1996.

[4] Priyanka Jain, Balbir Kumar and Shail Bala Jain. “Discrete sine transform and its inverse - realization through recursive algorithms”, International Journal of Circuit Theory and Applications; Volume 36 Issue4.

[5] Xueyao Li, Hua Xie, and Bailing Cheng, “Noisy Speech Enhancement Based on Discrete Sine Transform”, 2006 First International Multi-Symposiums on Computer and Computational Sciences, ISBN: 0-7695-2581-4.

[6] Ing Yann Soon , Soo Ngee Koh , Chai Kiat Yeo, “Noisy speech enhancement using discrete cosine transform”, Speech Communication 24 (1998) 249-257.

[5] John G. Proakis, Dimitris G.Manolakis, “Digital signal processing: principles, algorithms and applications”, Prentice Hall, March 2002.

[6] Rabiner and Gold, “Theory And Applications Of Digital Signal Processing”, Prentice Hall, Inc., 1975

Picture Collage with Genetic Algorithm and Stereo vision

Hesam Ekhtiyar[1], Mahdi Sheida[2] and Mahmood Amintoosi[3]

[1] Faculty of Electrical and Computer Engineering, Sabzevar Tarbiat Moallem University, Sabzevar, Iran,

[2] Faculty of Electrical and Computer Engineering, Sabzevar Tarbiat Moallem University, Sabzevar, Iran,

[3] Faculty of Mathematics and Computer Science, Sabzevar Tarbiat Moallem University, Sabzevar, Iran,

Abstract

In this paper, a salient region extraction method for creating picture collage based on stereo vision is proposed. Picture collage is a kind of visual image summary to arrange all input images on a given canvas, allowing overlay, to maximize visible visual information. The salient regions of each image are firstly extracted and represented as a depth map. The output picture collage shows as many visible salient regions (without being overlaid by others) from all images as possible. A very efficient Genetic algorithm is used here for the optimization. The experimental results showed the superior performance of the proposed method.

Keywords: *Picture Collage, Image Summarization, Depth Map*

1. Introduction

Detection of interesting or ”salient” regions is a main sub-problem in the context of image tapestry and photo collage. An ideal image summary should contain as many informative regions as possible on a given space [1]. The image mosaic can be considered as a simple form of photo collage, in which the images are placed on a canvas, side by side, without any rotation or considering informative regions. Google’s Picasa overlays images without considering any salient regions.

Figure 1 shows the left images of a collection of stereo images used in this paper.

Figure 2 shows some collages produced by the aforementioned methods or software over the images shown in figure 1. Images are randomly placed on a canvas. In figure 2 even the whole of all images are visible or the images are occluded, the importance of regions are discarded.

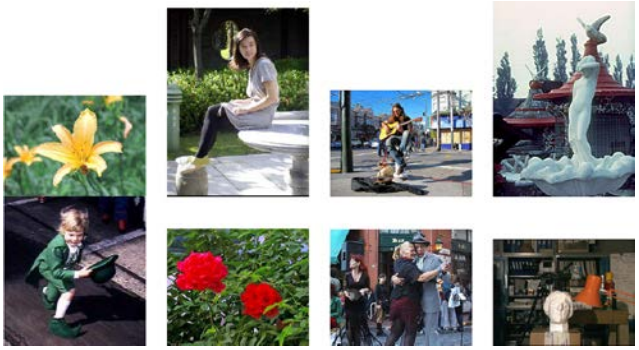


Fig. 1

LEFT IMAGES OF SOME STEREO IMAGES USED IN THIS PAPER.

An approach named saliency-based visual attention model [2] is used in [1] for extracting interesting regions. This model combines multi scale image features (color, texture, orientation) into a single topographical saliency map. In [3] three options are considered for saliency regions:

- A heuristic assumption that the image center

Is more informative about the image’s content than the border,

- it is assumed that blocks with high contrast are

Salient, and

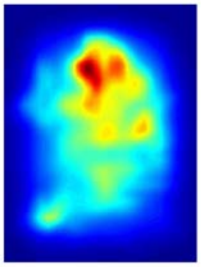
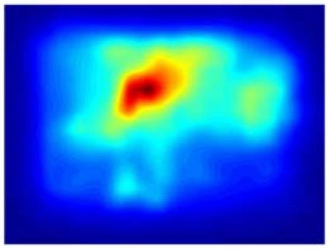
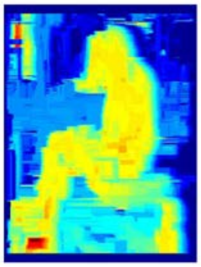
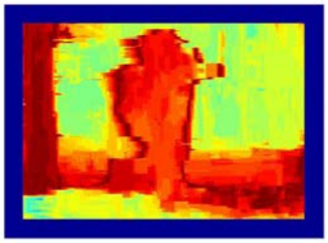
- it is assumed that typically tapestry is on a

Personal collection and a face detection module is used.

In recent years, capturing stereo images for various purposes such as 3D reconstruction has been usual. Here with this assumption that the focused object is the important part of the image and these parts are close to the camera, we used the depth map image of stereo image pairs for estimation of informative regions in a photo collage application.

The rest of the paper organized as follow: section 2 explains the proposed method. Section 3 provides the experimental results and section 4 is dedicated to the concluding remarks.

Table 1
SOME STEREO IMAGES WITH THEIR IMPORTANT REGIONS BASED ON SALIENCY MAP AND DEPTH MAP.

	Girl	Musician
Left Image		
Right Image		
Saliency Map		
Depth Map		

2. The Proposed Method

In order to obtain a pleasant photo collage the following properties should be considered [1], [4]:

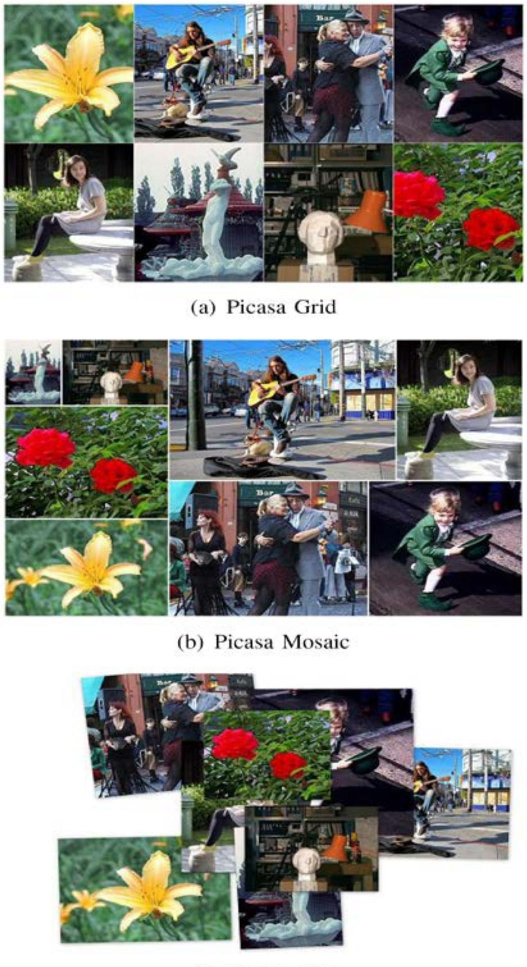
- Salience Maximization, to maximize the total amount of visible important regions as possible (A_{occ});
- Blank space minimization, to minimize the portions of the canvas which is not covered by any image (B),
- Salience Ratio Balance, for avoiding those cases in which a very small region of some images is visible (V).

The fitness function is a combination of the above three parameters: $\lambda_A.A_{occ} + \lambda_B.B + \lambda_V.V$.

Computing the depth map has been done here with a dynamic programming approach [5]. Table 1 shows two instance image pairs and their important regions with Saliency toolbox[2][1] and stereo depth map.

Saliency map is about to left image. As can be seen, the depth map image is more informative than saliency map.

The overall framework of the proposed method is shown in figure 3.



(a) Picasa Grid

(b) Picasa Mosaic

(c) Picasa Pile

Fig. 2
SOME COLLAGE OUTPUTS OF GOOGLE PICASSA.

[1] http://www.saliencytoolbox.net/

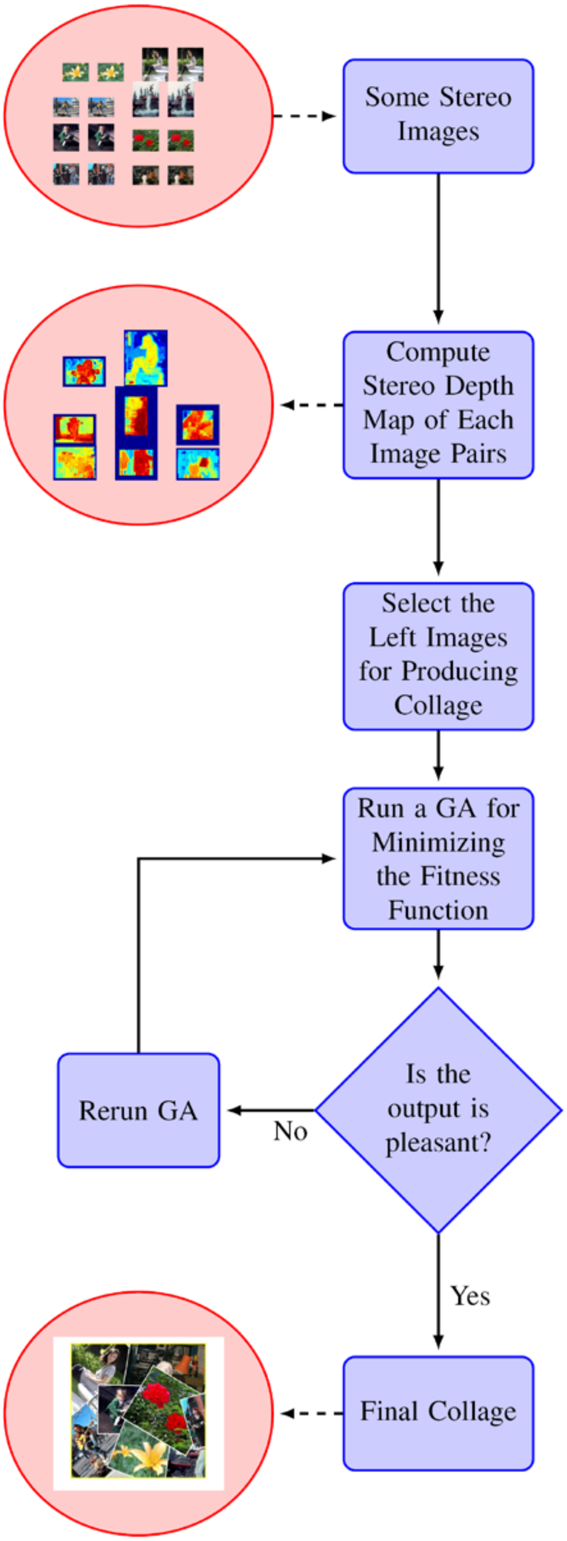

Some Stereo Images
Compute Stereo Depth Map of Each Image Pairs
Select the Left Images for Producing Collage
Run a GA for Minimizing the Fitness Function
Is the output is pleasant?
No
Rerun GA
Yes
Final Collage

Fig. 3
THE OVERALL FRAMEWORK OF THE PROPOSED METHOD.

3. Experimental Results

The implementation has been done using MATLAB 7. The population size of GA was 40, number of generations was 150. In the fitness function, we set the weights λA, λB and λV as 10/30, 9/30 and 11/30 respectively. The canvas is square and its size is set so that its area is about half of the total area of all input images. The input images (shown in figure 1) are gathered by searching over the Internet. Figure 4 show the value of GA fitness function over 150 generation.

Figure 5(a) shows the collage initiated at the first iteration of GA, 5(b) shows the final result of the proposed approach. Figure 5(c) shows depth map (as saliency region) occlusion, canvas usage, image rotations and cropping at the final stage of GA. As can be seen in figure 5(c) the most important regions of each image is visible in the final result.

Instead of the depth map image, we used saliency map for comparison purposes. Since the important regions with these two methods are different, the quantitative comparisons of these methods - via GA fitness function- is not possible. Hence we compared them by visual inspection of the produced collage with each method. Each method has been executed 20 times, and 2 of the best produced collages has been selected visually and illustrated in figure 6 . As can be seen the proposed method is competitive with the saliency map.

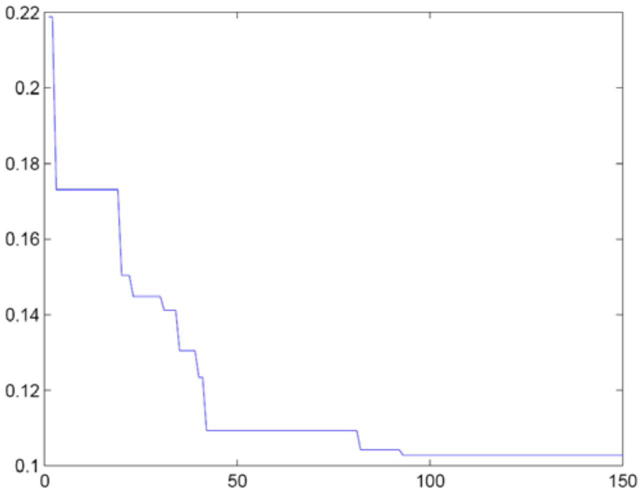
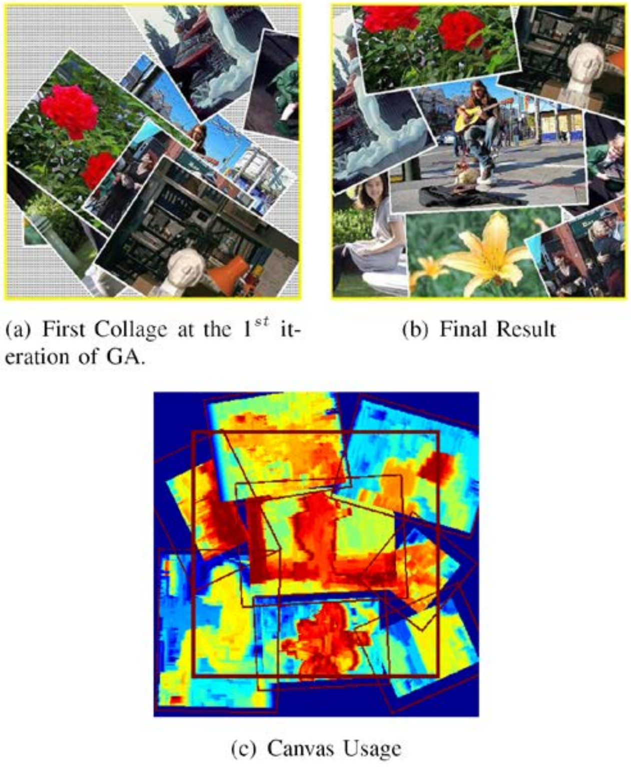


Fig. 4
THE VALUE OF FITNESS FUNCTION OVER GA ITERATIONS.

4. Conclusion

Until now in the photo collage context, only the mono images has been used; but here the stereo images were used for creating photo collage. In the previous works, the main source for determining important regions was based on: contrast, color, face detection and so on. Here with this assumption that usually the interested object is closer to the viewer with respect to other parts of the scene, the depth map of the stereo images has been used as an estimation of the image regions' importance.

The experimental results showed the good performance of the proposed method. Although in this paper only the stereo images has been used, but whenever some mono images and some stereo images are in hand, it is easy to use depth map for stereo images and another salient map extraction method for mono images. As future works we plan to implement the aforementioned idea and to extend our method to video collage.



(a) First Collage at the 1^{st} iteration of GA.

(b) Final Result

(c) Canvas Usage

Fig. 5
THE FIRST AND FINAL COLLAGE PRODUCED BY GA AT THE FIRST AND LAST ITERATIONS.

References

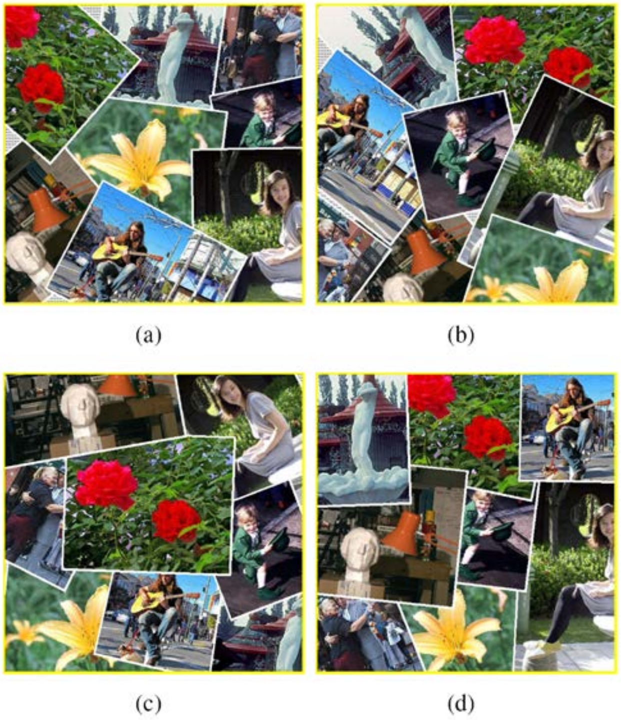
[1] J. Wang, J. Sun, L. Quan, X. Tang, and H.-Y. Shum, "Picture Collage," in CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), 2006, pp. 0–7.

[2] C. K. L. Itti and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Trans. Pattern Anal. Mach. Intell, vol. 20, no. 11, pp. 1254–1259, 1998.

[3] C. Rother, S. Kumar, V. Kolmogorov, and A. Blake, "Digital Tapestry," in CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1, 2005, pp. 589–596.

[4] S. Battiato, G. Ciocca, F. Gasparini, G. Puglisi, and R. Schettini, "Adaptive multimedial retrieval: Retrieval, user, and semantics," N. Boujemaa, M. Detyniecki, and A. N¨urnberger, Eds. Berlin, Heidelberg: Springer-Verlag, 2008, ch. Smart Photo Sticking, pp. 211–223.

[5] R. Szeliski, Computer Vision: Algorithms and Applications (Texts in Computer Science), 1st ed. Springer, Nov. 2010.



(a)

(b)

(c)

(d)

Fig. 6
THE RESULT OF THE GA, WHEN EACH OF THE SALIENCY MAP (A,B) OR DEPTH MAP (C,D) ARE USED FOR DETERMINING THE INFORMATIVE REGIONS.

Hesam Ekhtiyar received the B.S. degree in computer engineering from Sabzevar Tarbiat Moallem University, Sabzevar, iran, in 2011. his research interests include computer vision, speech recognition, robotics, soft computing.

Mahdi Sheida received the B.S. degree in computer engineering from Sabzevar Tarbiat Moallem University, Sabzevar, iran, in 2011. his research interests include computer vision, speech recognition, network programming.

Mahmood Amintoosi is an assistant professor in Sabzevar Tarbiat Moallem University. He received his B.Sc. degree in Mathematics and M.Sc. degree in Computer Engineering in 1994, 1998, respectively from Ferdowsi University of Mashhad. From 1998 to 2005, he was a Lecturer in the Department of Mathematics of Sabzevar Tarbiat Moallem University. He received his Ph.D. degree in Artificial Intelligence from Iran University of Science and Technology in 2011. His research interests include Computer Vision, Super-Resolution, Panorama, Automated Timetabling and Combinatorial Optimization. He has more than 30 conference and journal papers.

Multi agent Simulation: A Unified Framework for the analysis of viral infections within a bovine population

Tahar Guerram [1], Nour El Houda Dehimi [2]

[1]Département of Mathematics and computer Science, University of Oum El Bouaghi, Algeria

[2]Département of Mathematics and Computer Science, University of Oum El Bouaghi , Algeria

Abstract

Multi-agent systems [1] is an approach which allows to study population dynamics from a qualitative point of view by defining the attributes and the behaviors for the interacting individuals of the system . So, multi agent systems allow the design and the implementation of the individual models proposed for the study of population dynamics which have major advantages compared to the aggregate models. The present paper presents a multi agent based framework for the simulation of the impact of viral infections on a population of cows which may be modern (only females) or mixed. This simulation creates an artificial life [2] [3] which will make it possible to the user to feel in a virtual laboratory and will facilitate to him the forecasts of the impact of viral infections on the evolution of the targeted population..

Keywords: *multi agent systems, viral infection, modeling and simulation, population dynamics, artificial life.*

1 – Introduction

The dairy production estimated to two Billion liters per year covers only the 2/3 of the Algerian citizens needs. The remainder is a billion liters of milk, imported in the form of dried milk and represents an invoice exceeding a Billion of Dollars. This involves enormous economic losses and does not allow Algeria to have its food independence as regards dairy production. To cure in this irrefutable fact, Algeria tries to develop the bovine breeding, by importing of cows with high genetic potential. Unfortunately these an imals, too selected for the dairy production, became too sensitive to various pathologies, in particular the viral diseases. These last are often underhand and with fast trans mission, so that when the veterinary surgeons decide for a decision making, it is often too late. The ideal would be to have tools able to envisage the evolution of these diseases, and which make it possible to anticipate the fatal consequences of these viral diseases by fast and suitable decision making. The discounted objective of this work is the creation of thi s kind of tools. It is a question of proposing a unified framework allowing the study of the evolution of the viral infections within a population of dairy cows. The dynamics of the interactions between the indivi duals (cows) is a nonlinear dynamics (nonlinear evolution in time), which will encourage us to call upon the individual models of the dynamics of the pop ulations "the individual is the handled basic entity" [4]. In this tool each individual will be represented by an agent to which we assign a set of attributes and behaviors allowing us to follow its evolution [5]. The agents carry out their procedures simultaneously and are distinguished from/to each other; the addition or the withdrawal of an agent or a set of agents is easy [6], which will enable us to be closer to reality.

The proposed tool is able to study all the viral diseases within a high bovine population in a modern way or on a mixed population. The first type of population is only made up of females since the males are sold after 3 months as of their birth and then stockbreeders use artificial insemination for the reproduction of these cows. The second type of population is made up of males and females with various ages. To carry out simulations, the user of this tool will limit himself to introduce the parameters of the disease on which he wants to

make a study and the characteristics of the studied populati on (for example: the number of patients at the time of the declaration of the disease and the number of males in the population).

2. Related work

In the first work [7], the authors used the Multi agent system approach to model the impact of the Virus of the Human Immunodeficiency (HIV) on a population made up of cells intervening during the infection. In order to show the effectiveness of the individual model compared to the aggregate model based on mathematics. In this work the objective consists in creating a virtual environment in which the various agents evolve and interact between them. It is an environment with three dimensions corresponding to 1 mm^3 of blood. Three classes of agents were conceived simulating cells which are: The agents cells CD4, the agents infected cells CD4, the agents virus VIH besides another agent representing body THYMUS responsible for the production of cells CD4. The results obtained show the benefits of the Multi agent system approach making it possible to bring us closer to the reality.

The second work [8] aiming to study the impact of the practices of breeding and measurements of control on the dynamics of CIRCOVIRUS type-2 infection within a porcine population. In this work, a stochastic mathematical model in discrete time describing the dynamics of population in a porcine breeding of the standard borne -fattener, was developed. The results of this study show in particular the effect of the preventive measures, such as vaccination on the attenuation of the epidemic propagation.

The first work uses an individual model and implements it by a multi agent system but takes into account only one type of infection. In the second work, besides its taking into account of only one type of infection, rests on a discrete and stochastic model of the epidemic thus supposes automatically that the individuals have similar behaviors but each individual is primarily influenced by the behavior of the individuals of its entourage.

3. The Adopted approach

Contrary to similar work, our approach tries to propose a general framework of multi agent simulation taking into account a large range of viral infections within a population of cows. The discounted goal is to provide a powerful tool for simulation to be used by veterinary surgeons and epidemiologists enabling them to envisage the evolution of these diseases on a modern or mixed population and to prevent the consequences of these last by adequate decision making at the convenient period. We tried to approach to the maximum of reality by taking into account of two types of modern and traditional breeding, the consideration of the horizontal/vertical transmission of the diseases. Also, and by preoccupation with extensibility, the various parameters of the viruses are modifiable by the user because these factors can vary during the duration of study, for example increase in the virulence of the virus. Before describing our multi agent based approach, we will make a panorama on the domain

of discourse.

3.1 Description of the domain of discourse

The cows produce, in general a calf a year, they give their first calf when they are three years old. Their average lifetime is 15 years, during which, they can produce 10 calves. They live in certain promiscuity and have an average space of 10 m^2 by cow. The exploitation is only made up of females (dairy cows and heifers of replacement), the males are sold as of their birth. The stockbreeder uses artificial insemination for the reproduction of his cows. According to the objectives of the stockbreeders, we can find two types of populations: population only made up of females (dairy cows and heifers of replacement), the males are sold as of their birth. The stockbreeder uses artificial insemination for the reproduction of his cows (modern breeding). The second type of population is called “mixed”, i.e. composed of males and females at various ages.

During their diseases the animals can transmit microbes to other animals. This transmission of disease is called “contagion” and its rate varies according to the nature of disease. The sick animals can cure if they develop antibodies which confer immunity to them. The time necessary to obtain an immunity (lasted of immunity) depends on the type of the disease. The chances so that a sick animal acquires immunity depend on its conditions of breeding (food, hygiene, and parasitism). Indeed weak animals are unable to resist, they are more sensitive to the disease, and can die throughout installation of immunity (mortality rate during the acute phase). The more resistant animals, can resist the disease and develop immunity (chance of survival), some among can, nevertheless to die, but mortality rate in this case is weaker. Besides these parameters, it is necessary to take into account the fact that certain diseases are transmissible from mothers to calves during their birth.

3.2 Agents of our system and their interactions

Each agent is characterized by a set of attributes and of behaviors. The attributes selected are: Age, Disease_ Counter, Number_ of _ Children, sex (M/F), Immunized (O/N), Patient (O/N). The agents are distinguished by behaviors common to all the individuals: (Ageing, Displacement), behaviors common to animals of female sex (Reproduction) and behaviors of the sick animals: (Infected,Recovered). The interactions between the agents are expressed by the transfer of the disease (contagion) which is concretized in the behaviors Infected and Displacement. Figures 1 and 2 illustrate these interactions.

4. Implementation and results of simulation

We used Netlogo platform [9] which is a multi-agent programming language and an environment for the simulation of natural and social phenomena. It is particularly well adapted for the simulation of complex systems evolving through time.

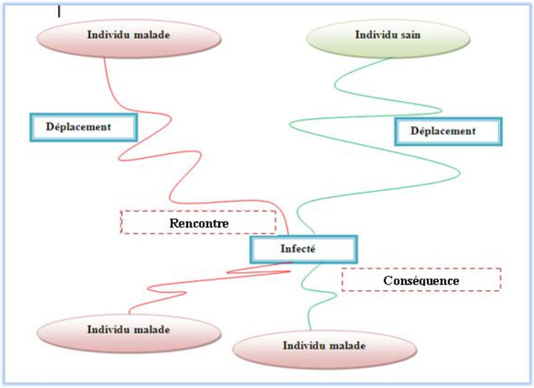


Fig. 1 Interaction between agents

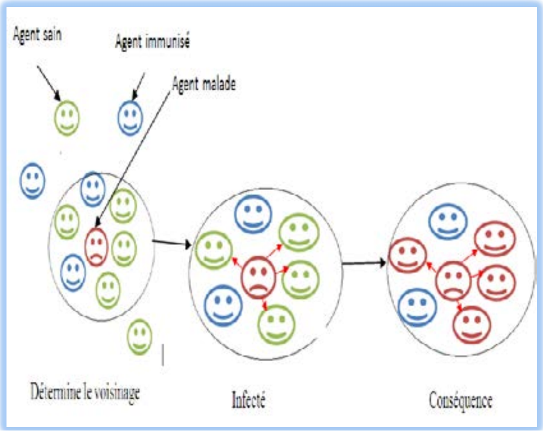
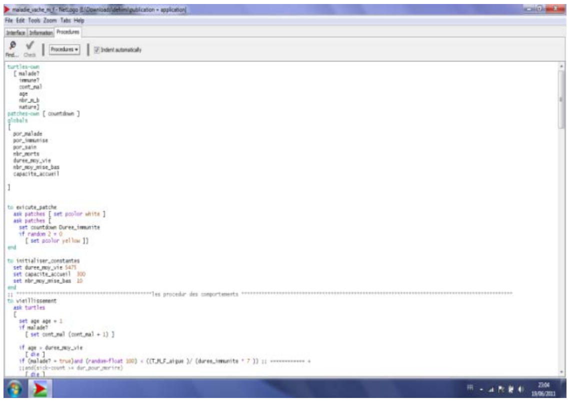


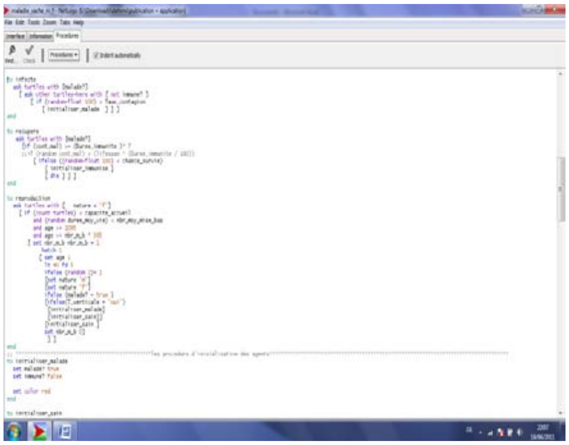
Fig. 2 Another representation of the Interaction between agents

Modelers can allot instructions to hundreds or thousands of the independent agents in order to make operations simultaneously. This is why, one can explore connections between small degree behaviors of agents and the great degree ones which emergent after their interactions. The world of Netlogo consists of reactive and mobile agents which can carry out behaviors in a simultaneous way. Three types of agents exist under Netlogo: *Patches*, *Turtles* and *Observer*. Turtles represent agents. The patch represents the environment of the agents . The evolution of simulation is managed by the Observer agent. Figure 3 and figure 4 show part of the developed code. With an aim of illustrating the interest of our application, we will take as example certain typical viral diseases (See Table 1). We will base ourselves on the specific parameters which characterize each disease namely: rates of contagion, of mortality in acute phase, duration of immunity acquisition of and the chance of survival and we will establish a forecast of the impact of these diseases on the evolution of the population namely: patient, healthy, immunized (Figures 5,6; Figures 7,8; Figures 9,10; Figures 11 ,12).


maladie_vache_m_f - NetLogo (E:\Download\defim\publication + application)
File Edit Tools Zoom Tabs Help
Interface Information Procedures
Find... Check Procedures ▾ ☑ Indent automatically
23:04
19/06/2011

```
turtles-own
  [ malade?
    immune?
    cont_mal
    age
    nbr_m_b
    nature]
patches-own [ countdown ]
globals
[
  por_malade
  por_immunise
  por_sain
  nbr_morts
  duree_moy_vie
  nbr_moy_mise_bas
  capacite_accueil

]


to exicute_patche
  ask patches [ set pcolor white ]
  ask patches [
    set countdown Duree_immunite
    if random 2 = 0
      [ set pcolor yellow ]]
end

to initialiser_constantes
  set duree_moy_vie 5475
  set capacite_accueil  300
  set nbr_moy_mise_bas  10
end
;; ********************************************les procedur des comportements ********************************************************************************************
to vieillissement
  ask turtles
  [
    set age age + 1
    if malade?
      [ set cont_mal (cont_mal + 1) ]

    if age > duree_moy_vie
      [ die ]
    if (malade? = true)and (random-float 100) < ((T_M_F_aigue )/ (duree_immunite * 7 )) ;; ************ +
    ;;and(sick-count >= dur_pour_morire)
      [ die ]
```

Fig. 3 A part of the developed code


File Edit Tools Zoom Tabs Help
Interface Information Procedures
Find... Check
Procedures
☑ Indent automatically

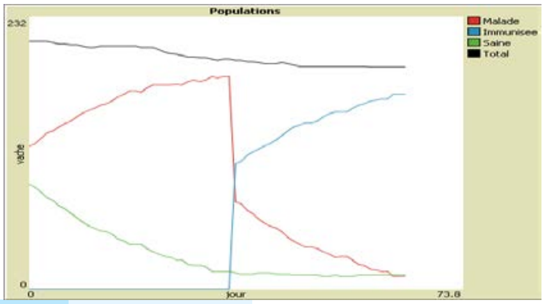
```
to infecte
  ask turtles with [malade?]
    [ ask other turtles-here with [ not immune? ]
       [ if (random-float 100) < taux_contagion
            [ initialiser_malade  ] ] ]
end

to recupere
   ask turtles with [malade?]
    [if (cont_mal) >= (duree_immunite )* 7
    ;;if (random cont_mal) > ([illegible] * (duree_immunite / 100))
        [ ifelse ((random-float 100) < chance_survie)
             [ initialiser_immunise ]
             [ die ] ] ]
end

to reproduction
  ask turtles with [   nature = "F"]
   [ if (count turtles) < capacite_actuel
        and (random duree_moy_vie) < nbr_moy_mise_bas
        and age >= 1095
        and age >= nbr_m_b * 365
     [ set nbr_m_b nbr_m_b + 1
          hatch 1
         [ set age 1
           [illegible]
           ifelse (random 2)= 1
           [set nature "M"]
           [set nature "F"]
           ifelse (malade? = true )
           [ifelse(T_verticale = "oui")
            [initialiser_malade]
            [initialiser_sain]]
           [initialiser_sain ]
           set nbr_m_b 0]
            ]]
end
;; *********************************Les procedures d'initialisation des agents*********************************
to initialiser_malade
  set malade? true
  set immune? false

  set color red
end

to initialiser_sain
```

Fig. 4 Another part of the developed code

Table 1: parameters of certain typical Viral diseases

Disease	incubation	Contagion rate	Mortality rate	cause	Chance of survival	Installation of immunity	Contagion mode	transmission vertical
Foot-and-mouth disease	2 - 7 days	80%	5%	virus	99%	21 days	contact	no
Plague of the ruminants	3 – 10 days	90%	80%	virus	90%	10 days	contact	no
Infectious Rhinotracheine (I.B.R)	2 - 4 days	80%	15%	Virus	98%	35 days	contact	yes
Virus of the bovine viral diarrhea (B.V.D)	07 days	85%	50%	virus	90%	15 days	contact	no

I.B. R



Plague of the ruminants

Fig. 5 Appearance of immunized individuals

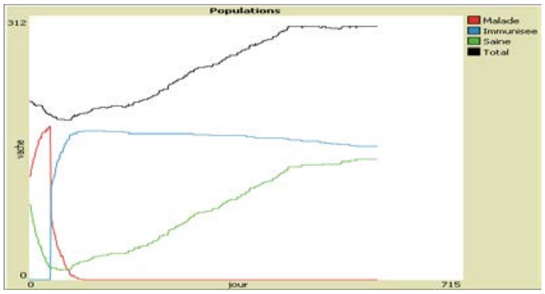


Fig. 6 Stabilization

B.V.D

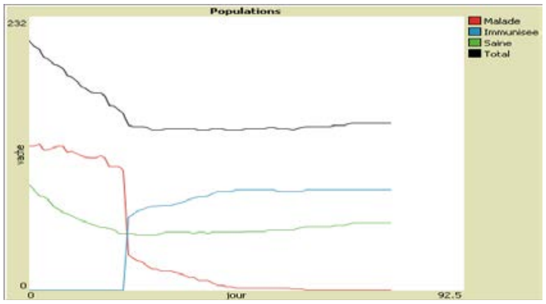


Fig. 7 Appearance of immunized individuals

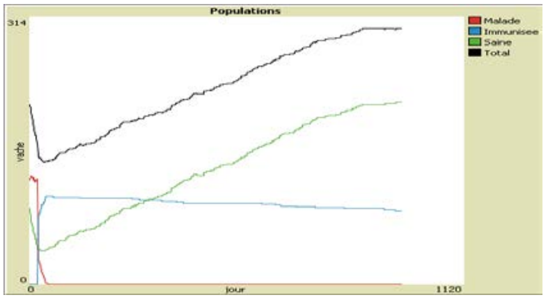


Fig. 8 Stabilization

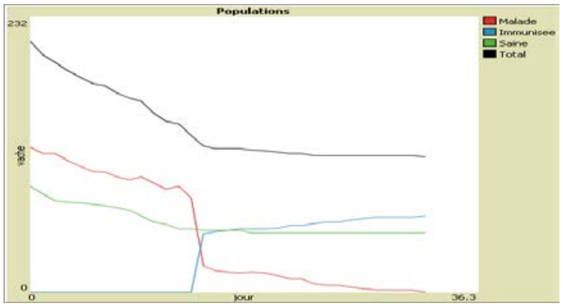


Fig. 9 Appearance of immunized individuals

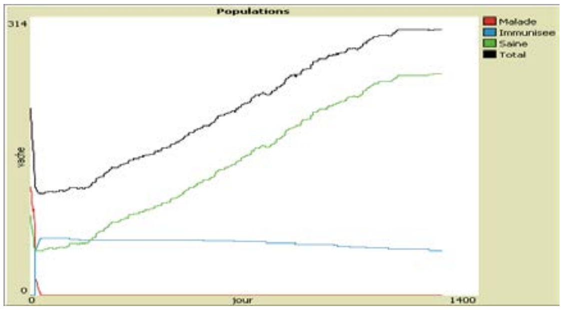


Fig. 10 Stabilization

Foot-and-mouth disease

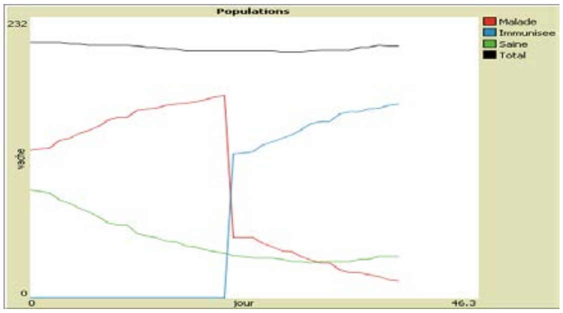


Fig. 11 Appearance of immunized individuals

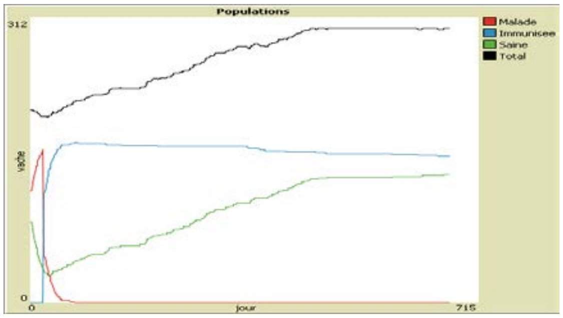


Fig. 12 Stabilization

5. Conclusion

Population dynamics behaves as a complex system with an unforeseeable evolution in time, the study of this kind of systems must be done using the individual directed model which justified our choice for the use of agent based modeling and simulation approach. Indeed, the use of the agent paradigm enabled us to represent the individuals as agents, distinguished from each other, each one being able to move, reproduce and transmit the disease. Also, it enabled us to produce a tool able to envisage the impact of the viral diseases in a reliable and perennial way, very useful for the epidemiologists, offering a practical complement to them to their biological theories in the fight against the underhand and unforeseeable effects of the viral diseases. The only disadvantage would be the fact that the simulation of the typical cases takes much time but on the other hand gives better results. Nevertheless, our work remains open to other improvements namely:

- Generalizing our tool so that it will be able to deal with any type of population.
- Introducing fuzzy parameters enabling us to create classes of cases of simulation thus facilitating decision making for the users of this tool.

References

[1] J. Ferber, Multi-Agent Systems: An Introduction to Distributed Artificial Intelligence. Addison-Wesley Longman Publishing Co., Inc., 1999.

[2] D. Floreano and S. Nolfi , “Adaptive behavior in competing co-evolving species”, In the Proceedings of the 4th European Conference on Artificial Life , 1997, Edited by Husbands P., Harvey I., MIT Press, Cambridge, p.378–387.

[3] J. Ventrella., “Emergent morphology and locomotion without a fitness function”, In the Proceedings of the 4th International Conference on Simulation of Adaptive Behavior, 1996, Edited by Maes P., Mataric M.J., Meyer J.A. PollackJ., Wilson S.W., MIT

Press, p.484–495.

[4] A. Lomnicki, “Population ecology from the individual perspective". In Individual-based models and approaches in ecology, 1999, Edited by de Angelis D. and Gross L. J. Chapman and Hall, New York, p.3–17.

[5] M. Bouzid,. “Contribution à la modélisation de l’interaction Agent/Environnement, modélisation stochastique et simulation parallèle », Thèse de doctorat de l’université Henri Poincaré, Laboratoire Lorrain de recherche en informatique et ses applications, Nancy, France, 2001.

[6] P. Ballet, “Intérêts Mutuels des Systèmes Multi-agents et de l’Immunologie. Applications à l’immunologie, l’hématologie et au traitement d’image“, Thèse de doctorat Université De Bretagne Occidentale, France, 2000.

[7] L. Toufik, and T. Bornia. “Modélisation en dynamique des populations, intérêt de l’approche Multi-Agents“ , in Proceedings of MajecSTIC, 2009, Avignon, France, 16- 18 Novembre 2009.

[8] A. Mathieu, G. Béatrice, A. J, François, and R. Nicolas, “Étude de l’impact des pratiques d’élevage et de mesures de maîtrise sur la dynamique d’infection par le Circovirus porcin de type 2: une approche par modélisation“, Bulletin Epidémiologique N° 33, 2009, pp : 7-10 http://www.afssa.fr/bulletin-epidemiologique/Documents/BEP-mg-BE33.pdf

[9] NetLogo Uri Wilensky. NetLogo 4.0.2 User Manual. http://ccl.northwestern.edu/netlogo/docs/NetLogoUser Manual.pdf

Tahar Guerram: Dr. Tahar Guerram is a lecturer of Computer Science at the Department of Mathematics and Computer Science of the University of Oum El Bouaghi in Algeria. His main areas of interest include qualitative and uncertain reasoning and simulation in mulli agent systems.

Nour El Houda Dehimi : Miss. Nour El Houda Dehimi is a Master degree student at the Department of Mathematics and Computer Science of the University of Oum El Bouaghi in Algeria. Her area of interest is Artificial Intelligence and the simulation of population dynamics.

A Comparative Study of Multi-Hop Wireless Ad-Hoc Network Routing Protocols in MANET

Tamilarasan-Santhamurthy

Department of Information Technology, LITAM,
Dullipala (village),Sattenpalli (Mandal),Guntur, Andhra Pradesh,522412,India

Abstract

Mobile Ad-Hoc Network (MANET) is a collection of wireless mobile hosts forming a temporary network without the aid of any stand-alone infrastructure or centralized administration. Most of the proposed MANET protocols do not address security issues. In MANETs routing algorithm is necessary to find specific routes between source and destination. The primary goal of any ad-hoc network routing protocol is to meet the challenges of the dynamically changing topology and establish an efficient route between any two nodes with minimum routing overhead and bandwidth consumption. The existing routing security is not enough for routing protocols. An ad-hoc network environment introduces new challenges that are not present in fixed networks. A several protocols are introduced for improving the routing mechanism to find route between any source and destination host across the network. In this paper present a logical survey on routing protocols and compare the performance of AODV, DSR and TORA.

Keywords: *DSR, AODV, TORA, MANET.*

1. Introduction

Wireless networking is an emerging technology that allows users to access information and services electronically, regardless of their geographic position. Wireless networks have become increasingly popular in the computing industry. The applications of the ad hoc networks are vast [9]. Mobile Ad hoc network (MANET) is a self-organized network because it is an infrastructure less feature of networks. MANET is a collection of nodes. Each node can connect by wireless communication links, without any fixed station such as base station. In MANET each node can act as a router and connectivity is achieved in the form of multihop graph between the nodes [8].

A routing is a core problem in network for sending data from one node to another. Several protocols have been developed under the authority of Mobile Ad hoc networking group. MANET is a charter of Internet Engineering Task Force (IETF). Lots of research has also been done about the performance of ad hoc networks under varying scenarios. Different kind of metrics or Characteristics may be used to analyze the performance of an ad hoc network [7, 9].

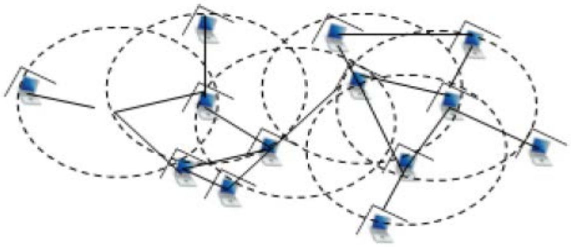


Fig. 1 Wireless Network Structures (Infrastructure less Networks)

1.1 Characteristics of MANET

- Dynamic Topology*:*
  Nodes can move arbitrarily with respect to other nodes in the network.
- Bandwidth-Constrained:
  MANET's nodes are mobile, so they are using radio links that have far lower capacity than hardwired link could use. In practice the realized throughput of a wireless network is less than a radio's theoretical maximum rate.
- Energy Constrained Operation:
  Mobile nodes are likely to relay on batteries, that is why the primary design criteria may sometimes be energy conservation.
- Limited Physical Security:
  Normally, radio networks are vulnerable to physical security threats compared to fixed networks. The possibility of eavesdropping, spoofing and Denial of Service attacks is higher. Existing link security techniques can be applied. However, a single point failure in an ad hoc network is not as crucial as in more centralised networks.
- Unpredictable Link Properties:
  Wireless media is very unpredictable. Packet collision is intrinsic to wireless network. Signal propagation faces difficulties such as signal fading, interference and multi-path cancellation. All these properties make the measures, such as

bandwidth and delay of a wireless link, unpredictable.

- Hidden and Exposed Terminal Problems:
  In the MAC layer with the traditional carrier sense multiple access (CSMA) protocol, multi-hop packet relaying introduces the “hidden terminal” and “exposed terminal” problems. The hidden terminal problem happens when signals of two, say B and C, which are out of the transmission range of each other, collide at a common receiver, say node A. An exposed terminal is created when a node A, is within range of and between two other nodes B and C, which are out of range of each other. When A wants to transmit to one of them, node B for example, the other node, C in this case, is still able to transmit to a fourth node, D which is in C’s range (but out of the range of node A). Here A is an exposed terminal to C but can still transmit to B.

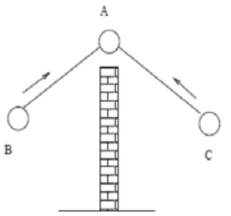


Fig.2: Hidden Terminal Problem

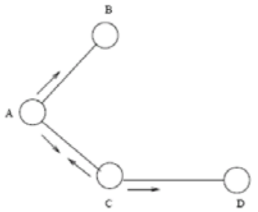


Fig.3: Exposed Terminal Problem

- Route Maintenance:
  The dynamic nature of the network topology and the changing behavior of the communication medium make the precise maintenance of network state information very difficult. Thus the routing algorithms in ad hoc networks have to operate with inherently imprecise information. Furthermore, in ad hoc networking environments, nodes can join or leave anytime. The established routing paths may be broken even during the process of data transfer. So, need for maintenance and reconstruction of routing paths with minimal overhead and delay.
  QoS-aware routing would require reservation of resources at the routers (intermediate nodes). However, with the changes in topology the intermediate nodes also change and new paths are created. Thus the reservation maintenance with the updates in the routing path becomes cumbersome.

1.2 Issues in MANETs:

- Multicasting:
  This is the ability to send packets to multiple nodes at once. This is similar to broadcasting except the fact that the broadcasting is done to all the nodes in the network. This is important as it takes less time to transfer data to multiple nodes.
- Loop Free:
  A path taken by a packet never transits the same intermediate node twice before it arrives at the destination. To improve the overall, we want the routing protocol to guarantee that the routes supplied are loop-free. This avoids any waste of bandwidth or CPU consumption.
- Multiple routes:
  If one route gets broken due to some disaster, then the data could be sent through some other route. Thus the protocol should allow creating multiple routes.
- Distributed Operation:
  The protocol should of course be distributed. It should not be dependent on a centralized node.
- *Reactive:*
  It means that the routes are discovered between a source and destination only when the need arises to send data. Some protocols are reactive while others are proactive which means that the route is discovered to various nodes without waiting for the need.
- Unidirectional Link Support:
  The radio environment can cause the formation of unidirectional links. Utilization of these links and not only the bi-directional links improves the routing protocol performance.
- Power Conservation:
  The nodes in an ad-hoc network can be laptops and thin clients, such as PDAs that are very limited in battery power and therefore use some sort of stand-by mode to save power. It is therefore important that the routing protocol has support for these sleep-modes.
- Proactive Operation:
  This is opposite to demand based operation. If additional delays that occur in demand based operation are unacceptable, proactive approach can be used especially when energy and bandwidth capacities support the use of proactive operation.
- Security:
  Ad hoc routing protocols are exposed too much kind of attacks. Maintaining link layer security is in practice harder with ad hoc networks than with fixed networks. Sufficient routing protocols security is desirable. Sufficient within

this context covers prohibiting disruption or modification of protocol operation.

- "Sleep" Period Operation:
  Since nodes in ad hoc networks may have energy constraints or because of some other need, nodes may want to stop sending and/or receiving data from arbitrary time periods. A routing protocol should be able handle such "sleep" periods without overly unfavourable consequences.

1.3 Applications of MANET:

- Sensor Networks for environmental monitoring.
- Rescue operations in remote areas.
- Remote construction sites and Personal Area Networking.
- Emergency operations.
- Military battlefield.
- Civilian environments.
- Law enforcement activities.
- Commercial projects.
- Educational Class rooms.

2. MANET Routing Protocol

MANET protocols are used to create routes between multiple nodes in mobile ad-hoc networks. IETF (Internet Engineering Task Force) MANET working group is responsible to analyze the problems in the ad-hoc networks and to observe their performance [7, 9]. There are different criteria for designing and classifying routing protocols for wireless ad-hoc networks. The MANET protocols are classified into three huge groups, namely Proactive (Table-Driven), Reactive (On-Demand) routing protocol and hybrid routing protocols [1, 2]. The following figure shows the classification of protocols.

Proactive (Table-Driven) routing protocol: - In proactive routing protocol perform consistent and up-to-date routing information to all the nodes is maintained at each node.

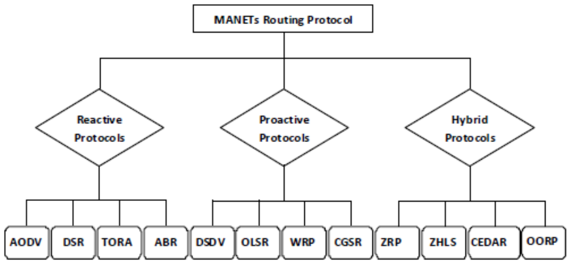


Fig.4 Different type of routing protocols in wireless Ad-hoc network

Reactive (On-Demand) routing protocol: - This type of protocols find route on demand by flooding the network with Route Request packets

2.1. Proactive vs. Reactive Routing

Proactive Schemes determine the routes to various nodes in the network in advance, so that the route is already present whenever needed. Route Discovery overheads are larger in such schemes as one has to discover all routes. Examples of such schemes are the conventional routing schemes, Destination Sequenced Distance Vector (DSDV). Reactive Schemes determine the route when needed. Therefore they have smaller Route Discovery overheads. Examples for such schemes are Ad Hoc On-Demand Distance Vector (AODV) routing protocol.

2.2. Single-Path vs. Multi-Path

There are several criteria for comparing single-path routing and multi-path routing in ad-hoc networks. First, the overhead of route discovery in multi-path routing is much more than that of single-path routing. On the other hand, the frequency of route discovery is much less in a network which uses multi-path routing, since the system can still operate even if one or a few of the multiple paths between a source and a destination fail. Second, it is commonly believed that using multi-path routing results in a higher throughput. Third, multi-path networks are fault tolerant when dynamic routing is used, and some routing protocols, such as OSPF (Open Shortest Path First), can balance the load of network traffic across multiple paths with the same metric value [2, 6, 10].

2.3. Proactive vs. Source Initiated

A proactive (Table-Driven) routing protocols are maintaining up-to-date information of both source and destination nodes. It is not only maintained a single node's information, it can maintain information of each and every nodes across the network. The changes in network topology are then propagated in the entire network by means of updates. Some protocols are used to discover routes when they have demands for data transmission between any source nodes to any destination nodes in network, such protocol as DSDV(.Destination Sequenced Distance Vector) routing protocol. These processes are called initiated on-demand routing. Examples include DSR (Dynamic Source Routing) and AODV (Ad-hoc On Demand Distance Vector) routing protocols [2].

3. AD-HOC on Demand Vector Protocol (AODV)

AODV combines some properties of both DSR and DSDV. It uses route discovery process to cope with routes *on-demand* basis. It uses routing tables for maintaining route information. It is reactive protocol; it doesn't need to maintain routes to nodes that are not communicating. AODV handles route discovery process with *Route Request (RREQ)* messages. *RREQ* message is broadcasted to neighbour nodes. The message floods through the network until the desired destination or a node knowing fresh route is reached. Sequence numbers are used to guarantee *loop freedom*. *RREQ* message cause bypassed node to allocate route table entries for reverse route. The destination node unicasts a *Route Reply (RREP)* back to the source node. Node transmitting a *RREP* message creates routing table entries for forward route [14].

For route maintenance nodes periodically send *HELLO* messages to neighbour nodes. If a node fails to receive three consecutive *HELLO* messages from a neighbour, it concludes that link to that specific node is down. A node that detects a broken link sends a *Route Error (RERR)* message to any upstream node. When a node receives a *RERR* message it will indicate a new source discovery process [5, 14].

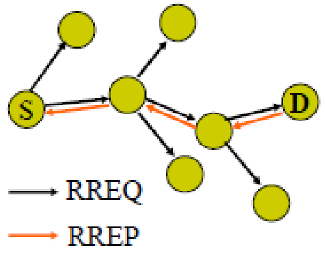


Fig. 5 AODV routing protocol with RREQ and RREP message

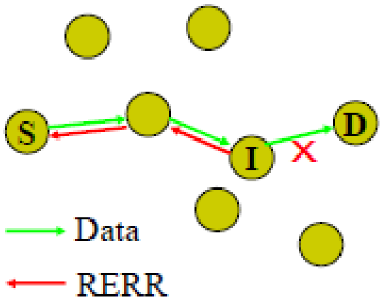


Fig 6 AODV routing protocol with RERR message

4. Dynamic Source Routing (DSR)

The Dynamic Source Routing Protocol (DSR) is a reactive routing protocol .By the means of this protocol each node can discover dynamically a source route to any destination in the network over multiple hops. It is trivially loop free owing to the fact that a complete, ordered list of the nodes through which the packet must pass is included in each packet header. The two main mechanisms of DSR are Route Discovery and Route Maintenance, which work together to discover and maintain source routes to arbitrary destinations in the network [1, 5]. The following figure shows the route discovery method.

Salvaging: An intermediate node can use an alternate route from its own cache, when a data packet meets a failed link on its source rout e.

Gratuitous route repair: A source node receiving a RERR packet piggybacks the RERR in the following RREQ. This helps clean up the caches of other nodes in the network that may have the failed link in one of the cached source routes.

Promiscuous listening: When a node overhears a packet not addressed to it, it checks if the packet could be routed via itself to gain a shorter route. If so the node sends a gratuitous RREP to the source of the route with this new, better route. Aside from this, promiscuous listening helps a node to learn different routes without directly participating in the routing process [5, 6].

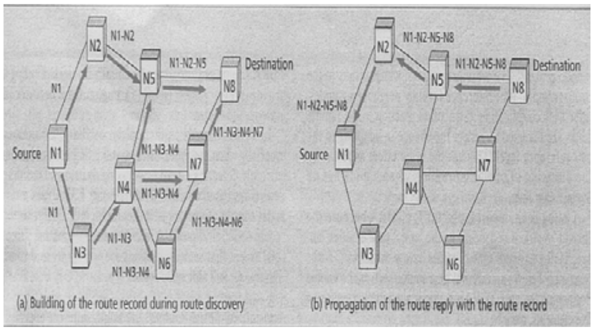


Fig.7 Creation of the route record in DSR

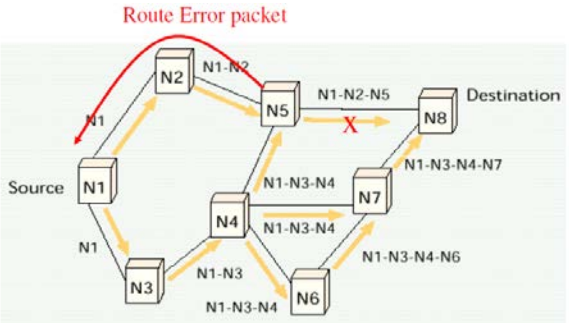


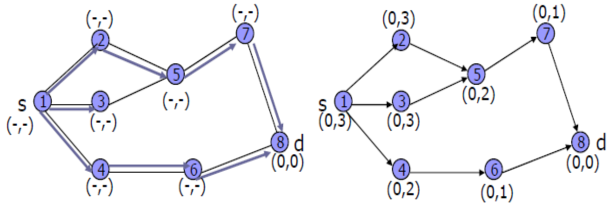
Fig. 8 Building of the route record during route discovery

5. Temporary Ordered Routing Algorithm (TORA)

The Temporary Ordered Routing Algorithm (TORA) is a highly adaptive, efficient and scalable distributed routing algorithm based on the concept of link reversal. TORA is proposed for highly dynamic mobile, multi-hop wireless networks. It is a source-initiated on-demand routing protocol. TORA finds multiple routes between source node and destination node. The main feature of TORA is that the control messages are localized to a very small set of nodes near the occurrence of a topological change. To achieve this, the nodes maintain routing information about adjacent nodes. The protocol has three basic functions:

- Route Creation,
- Route Maintenance and
- Route Erasure.

TORA can suffer from unbounded worst-case convergence time for very stressful scenarios. TORA has a unique feature of maintaining multiple routes to the destination so that topological changes do not require any reaction at all. The protocol reacts only when all routes to the destination are lost. In the event of network partitions the protocol is able to detect a partition and erase all invalid routes [19, 20].



Propogation of QRY
(reference level, height)

Height of each node
updated by UPD

Fig. 9: Route Creation of TORA

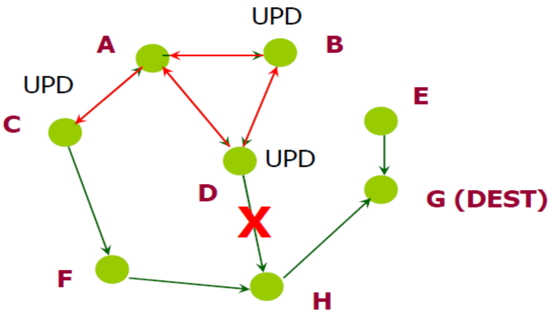


Fig. 10: Route Maintenance

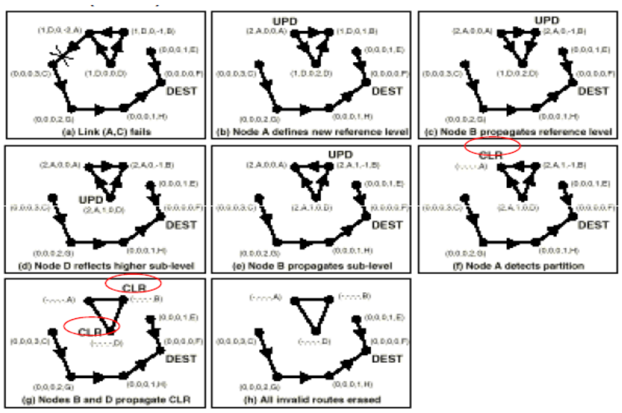

DEST
(a) Link (A,C) fails
UPD
(b) Node A defines new reference level
UPD
(c) Node B propagates reference level
UPD
(d) Node D reflects higher sub-level
UPD
(e) Node B propagates sub-level
CLR
(f) Node A detects partition
CLR
CLR
(g) Nodes B and D propagate CLR
(h) All invalid routes erased

Fig. 11: Erase Invalid Routes after a failure which Partitions the network

6. Simulation

The simulations were performed using Network Simulator 2 (Ns-2), particularly popular in the ad hoc networking community. The traffic sources are CBR (continuous bit –rate). The source-destination pairs are spread randomly over the network. The mobility model uses 'random waypoint model' in a rectangular filed of 500m x 500m with 50 nodes. During the simulation, each node starts its journey from a random spot to a random chosen destination. Once the destination is reached, the node takes a rest period of time in second and another random destination is chosen after that pause time. This process repeats throughout the simulation, causing continuous changes in the topology of the underlying network. Different network scenario for different number of nodes and pause times are generated [18].

Table 1: Simulation Parameters

SL.NO.	*PARAMETER*	*VALUE*
1.	Simulator	ns-2

2.	Protocols studied	AODV, DSR and TORA
3.	Simulation time	200 sec
4.	Simulation area	500×500
5.	Transmission range	250 m
6.	Node movement model	Random waypoint
7.	Bandwidth	2 Mbps
8.	Traffic type	CBR
9.	Data payload	Bytes/packet

7. Metrics for Performance Analysis

- Throughput: Ratio of the packets delivered to the total number of packets sent.
- Packet Delivery: Packet Delivery Ratio in this simulation is defined as the ratio between the number of packets sent by constant bit sources (CBR) and numbers of packets received by CBR sink at destination.

$$PktDelivery\% = \frac{\sum_{1}^{n} CBRrecv}{\sum_{1}^{n} CBRrecv} \times 100......Equation1$$

- Minimum Delay: Minimum Time taken for the packets to reach the next node.
- Maximum Delay: Maximum Time taken for the packets to reach the next node.
- Average End-to-End Delay: Time taken for the packets to reach the destination.

$$Avg\_End-to-End\_Delay = \frac{\sum_{1}^{n}(CBRsentTime - CBRrecvTime\_)}{\sum_{1}^{n} CBRrec}......Equation2$$

- Simulation Time: The time for which simulations will be run i.e. time between the starting of simulation and when the simulation ends.
- Network size: It determines the number of nodes and size of area that nodes are moving within. Network size basically determines the connectivity. Very lesser nodes in the same area mean fewer neighbours to send request to, but also smaller probability of collision.
- Number of Nodes: This is constant during the simulation. We used 50 nodes for simulations.
- Pause time: Node will stop a “pause time” amount before moving to another destination point.
- Jitter: Jitter describes standard deviation of packet delay between all nodes.
- Power Consumption: The total consumed energy divided by the number of delivered packet.
- Average Packet Delay: It is the sum of the times taken by the successful data packets to travel from their sources to destination divided by the total number of successful packet. The average packet delay is measured in seconds.
- Average Hop Count: It is sum of the times taken by the successful data packets to travel from their sources to destination divided by the total number of successful packets. The average hop count is measured in number of hops.
- Node Expiration time (NET)**:** it is the time for which a node has been alive before it must halt transmission due to battery reduction. The node expiration is plotted as number of nodes alive at a given time, for different point in time during the simulation.

8. Result Analysis

8.1. Packet Delivery Fraction (PDF) or Throughput

- TORA performs buffer at high mobility but in other cases it shows to have lower throughput.
- As per result AODV have the best overall performance.
- On-Demand protocols (DSR and DSDV) drop a considerable number of packets during the route discovery phase; a route acquisition takes time proportional to the distance between the source and destination.
- Packet drops are fewer with proactive protocols as alternate routing table entries can always be assigned in response to link failures.
- TORA can be quite sensitive to the loss of routing packets compared to the other protocols.
- Buffering of data packets while route discovery in progress, has a great potential of improving DSR, AODV and TORA performance.
- AODV has a slightly lower packet delivery performance than DSR because of higher drop rates.
- AODV uses route expiry, dropping some packets when a route expires and a new route must be

found

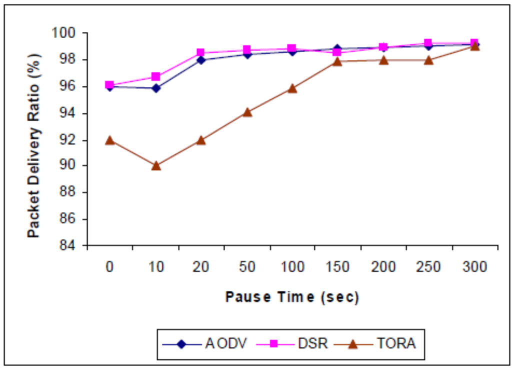


Fig.12: Packet delivery fraction vs. Pause time for 50-nodes with 10 sources.

8.2. End-to-End Delay

- AODV and DSR show poor delay characteristics as their routes are typically not the shortest over a period of time due to node mobility.
- AODV performs a little better delay-wise and can possibly do even better with some fine-tuning of this timeout period by making it a function of node mobility.
- TORA too has the worst delay characteristics because of the loss of distance information with progress.
- TORA route construction may not occur quickly.
- In DSR Route Discovery is fast, therefore shows a better delay performance than the other reactive protocols at low pause time (high mobility).
- In case of congestion (high traffic) DSR control messages get loss thus eliminating its advantage of fast establishing new route.
- Without any periodic hello messages, DSR outperforms the other protocols in terms of overhead.
- In most cases, both the packet overhead and the byte overhead of DSR are less than a quarter of AODV's overhead.
- The excellent routing load performance of DSR is due to the optimizations possible by virtue of source routing.
- TORA's performance is not very competitive with the distance vector and on-demand protocols.
- TORA shows a better performance for large networks with low mobility rate.

Sl.No	Protocol Property	AODV	DSR	TORA
1.	Multi-Cost Routes	NO	YES	YES
2.	Distributed	YES	YES	YES
3.	Unidirectional Link	NO	YES	YES
4.	Multicast	YES	NO	NO
5.	Periodic Broadcast	YES	NO	YES
6.	QoS Support	NO	NO	YES
7.	Routes Information Maintained in	Route Table	Route Cache	Adjacent Routers(One-Hop-Knowledge)
8.	Reactive	YES	YES	YES
9.	Provide Loop-Free Routers	YES	YES	YES
10	Route Optimization	YES	YES	YES
11.	Scalability	YES	YES	YES
12.	Route Reconfiguration	Erase Route Notify Source	Erase Route Notify Source	Link Reversed Route Repair
13.	Proactive	NO	NO	YES
14.	Routing Philosophy	FLAT	FLAT	FLAT

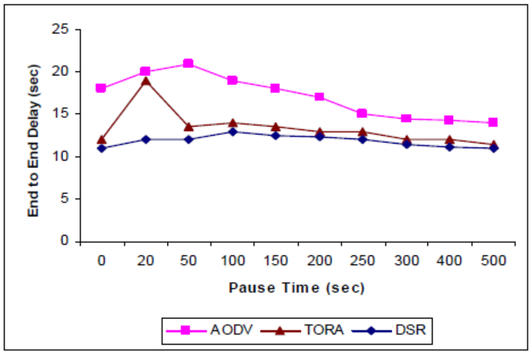


Fig.13: End- to -End Delay vs. Pause time for 50-node model with 10 sources.

Table 2: Comparison between AODV, DSR and TORA

9. Conclusion

This work is an attempt towards a comprehensive performance evaluation of three commonly used mobile ad hoc routing protocols (DSR, TORA and AODV). Over the past few years, new standards have been introduced to enhance the capabilities of ad hoc routing protocols. As a result, ad hoc networking has been receiving much attention from the wireless research community. In this paper, using the latest simulation environment NS 2, we evaluated the performance of three widely used ad hoc network routing protocols using packet-level simulation. The simulation characteristics used in this research, that is, packet delivery fraction and end-to-end delay are unique in nature, and are very important for detailed performance evaluation of any networking protocol. We can summarize our final conclusion from our experimental results as

- Increase in the density of nodes yields to an increase in the mean End-to-End delay.
- Increase in the pause time leads to a decrease in the mean End-to-End delay.
- Increase in the number of nodes will cause increase in the mean time for loop detection.

In short, AODV has the best all round performance. DSR is suitable for networks with moderate mobility rate. It has low overhead that makes it suitable for low bandwidth and low power network. TORA is suitable for operation in large mobile networks having dense population of nodes. The major benefit is its excellent support for multiple routes and multicasting.

References

[1] Sapna S.Kaushik; P.R.Deshmukh, (2009), Comparison of effectiveness of AODV, DSDV, and DSR routing protocols in MANETs. International Journal of Information Technology and Knowledge management," July-December 2009, Volume 2, No. 2, pp. 499-502.

[2] S. A. Ade; P.A.Tijare, (2010). "Performance Comparison of AODV, DSDV, OLSR and DSR Routing Protocols in Mobile Ad Hoc Networks." July-December 2010, Volume 2, No. 2, pp. 545-548.

[3] P.Johansson, T. Larsson, N. Hedman, B. Mielczarek, and M. Degermark. "Scenario based Performance Analysis of Routing Protocols for Mobile Ad-hoc Networks", Mobicom'99, 1999, Pages 195-206.

[4] C. E. Perkins and P. Bhagwat "Highly Dynamic Destination Sequenced Distance-vector Routing (DSDV) for Mobile Computers", Proceedings of the ACM SIGCOMM '94 Conference, August 1994, pages 234–244.

[5] S.R. Das, C.E. Perkins, and E.E. Royer, "Performance Comparison of Two on Demand Routing Protocols for Ad Hoc Networks," Proc. INFOCOM, 2000, pp. 3-12.

[6] Geetha Jayakumar and G. Gopinath; "Performance Comparison of Two On-demand Routing Protocols for Ad-hoc Networks based on Random Way Point Mobility Model" American Journal of Applied Sciences 5 (6): 2008, ISSN 1546-9239, pp. 659 – 664.

[7] R.M.Shrma; "Performance Comparison of AODV, DSR and AntHocNet Protocols", International Journal of Computer Science and Technology. Volume No. 1, Issue 1, September 2010, pp. 29 – 32.

[8] K. TAMIZARASU; M. RAJARAM; "Analysis of AODV Routing Protocol for Minimized Routing Delay in Ad Hoc Networks", International Journal on Computer Science and Engineering. Vol. 3, No. 3, March 2011, pp. 1075 – 1078.

[9] Md. Anisur Rahman; Md. Shohidul Islam; Alex Talevski; "Performance Measurement of Various Routing Protocols in Ad-hoc Network", Proceedings of International MultiConference of Engineers and Scientists, 2009, Vol. 1, pp. 18 – 20.

[10] Nor Surayati Mohamad Usop; Azizol Abdullah; Ahmad Faisal Amri Abidin; "Performance Evaluation of AODV, DSDV & DSR Routing Protocol in Grid Environment", IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.7, July 2009, pp. 261 – 268.

[11] Boukerche A, "Performance Comparison and Analysis of Ad Hoc Routing Algorithms", IEEE International Conference on Performance, Computing and Communications, 2001, Apr 2001, pp 171-178.

[12] G. Karthiga; J.Benitha Christinal; Jeban Chandir Moses; "Performance Analysis of Various Ad-Hoc Routing Protocols in Multicast Environment", International Journal of Computer Science and Technology. VO l. 2, Issue 1, March 2011, pp. 161 – 165.

[13] Samir R. Das; Charles E. Perkins; Elizabeth M. Royer; "Performance Comparison of Two On-Demand Routing Protocols for Ad Hoc Networks", IEEE Personal Communications Magazine special issue on Ad hoc Networking, February 2001, pp. 16-28.

[14] C.E.Perkin; Charles E. Perkins; Elizabeth M. Royer; " Ad hoc on-demand distance vector (AODV) routing", Proceedings of the 2nd IEEE Workshop on Mobile Computing Systems and Applications, New Orleans, LA, February 1999, pp. 90-100.

[15] R.Asokan; A.M.Natarajan; C.Venkatesh; "Optimized Quality of Service (QoS) Routing in mobile ad hoc networks using SELF – HEALING Technique ", International Journal on Wireless & Optical Communications (IJWOC),2007. Vol. 4, Issue 3. pp. 291-304.

[16] Dharmaraju, D; Roy-Chowdhury, A; Hovareshti, P; Baras, J.S; "INORA – a unified signaling and routing mechanism for QoS support in mobile ad hoc networks", parallel Processing workshop, 2002, proceedings. International Conference on 2002, pp. 86-93

[17] Rakesh Kumar Jha; Suresh V. Limkar; Dr. Upena D. Dalal; "A Performance Comparison of Routing Protocols (DSR and TORA) for Security Issue In MANET(Mobile Ad Hoc Networks)", IJCA Special Issue on "Mobile Ad-hoc Networks" MANETs, 2010, pp. 79-83.

[18]. NS-2 Network simulator http://www.isi.edu/nsnam/ns.

[19]. Park V. and S. Corson, 2001. Temporary-ordered Routing Algorithm (TORA). Internet Draft, draft-ietf-manettora-spec-04.txt.

[20]. V. Park and S. Corson, Temporally Ordered Routing Algorithm (TORA) Version 1, Functional specification IETF Internet draft (1998), http://www.ietf.org/internet-drafts/draft-ietf-manet-tora-spec-01.txt.

[21]. Samir R. Das, Robert Castaneda and Jiangtao Yan, "Simulation based performance evaluation of routing protocols for mobile ad hoc networks".

About the author

S. Tamilarasan, M.E.
Associate professor cum Head of Department, Loyola institute of Technology and management,
Dullipala (village), Sattenapalli (Mandal), Guntur, Andhra Pradesh, India.
Specialization:
Mobile computing, Advanced Data Structure, Design and Analysis of algorithm, Computer networks

Named Entity Identifier for Malayalam Using Linguistic Principles Employing Statistical Methods

Bindu.M.S[1] and Sumam Mary Idicula[2]

[1] Dept.of Computer Science ,M.G University
Edappally, Cochin, Kerala, India

[2] Dept. of Computer Science, Cochin University of Science and Technology
Cochin, Kerala ,India

Abstract

Natural language processing (NLP) began as a branch of Artificial Intelligence is a field of computer science and linguistics and is concerned with interaction between human language and computer. Major tasks of NLP such as Machine Translation (MT), Information Retrieval (IR) and Summarization require extensive knowledge of the language for the effective identification of semantic information in the text. Meaning or semantics of a text is mainly decided by the named entities which are the role carrying agents in a text. The system presented here is a Named Entity (NE) Identifier created using Statistical methods based on linguistic grammar principles. Malayalam NER is a difficult task as each word of named entity has no specific feature such as Capitalization feature in English. NERs in other languages are not suitable for Malayalam language since its morphology, syntax and lexical semantics is different from them. For testing this system, documents from well known Malayalam news papers and magazines containing passages from five different fields are selected. Experimental results show that the average precision recall and F-measure values are 85.52%, 86.32% and 85.61% respectively.

Keywords: *Malayalam compound word, Finite state Transducer, Extended Conditional Random Field, Feature vector*.

1. Introduction

NER is an important tool in almost all natural language processing applications such as IE, IR and Question Answering (QA) systems. Proper identification and classification of NEs are very crucial and pose a big challenge to the NLP researchers. The level of ambiguity in NER makes it difficult to attain human performance [1].

NER is the process of identifying and categorizing names in text. The NE task was first introduced as part of the MUC 6 (MUC 1995) evaluation exercise and was continued in MUC 7(MUC 1998).This formulation of NE task defines 7 types of NE: PERSON, ORGANISATION, LOCATION, DATE, TIME, MONEY and PERCENTAGE. NER also known as entity identification and extraction is a subtask of IE that seeks to locate and classify atomic elements in text into predefined categories. In the expression named entity the word named restricts the task to those entities for which one or many rigid designators as defined by Kripke stands for the referent [2].

The term named entity is not strict but has to be explained in the context where it is used. Entity names form the main context of a document. NER is a very important step towards more intelligent IE and management. NER performs what is known as surface parsing ,delimiting sequences of tokens that answer important questions such as “what”, “where” and “how” in a sentence.

Malayalam belongs to the Dravidian family of languages and is one of the 4 major languages of this family. It is one of the 22 scheduled languages of India with official language status in the state of Kerala. It is spoken by 35.9 million people. Malayalam is a morphologically rich agglutinative language and relatively of free order. Also Malayalam has a productive morphology that allows the creation of complex words which are often highly ambiguous [3]. A lot of work has been done in the field of NER for English and European Languages. In English Capitalization is a major clue for identifying person names. Some efforts have been made for Telugu, Hindi and Bengali. As of now we have no information regarding Malayalam NER work and no tag set is been identified so far.

Conditional Random Field (CRF) is a probabilistic framework for labeling or segmenting data. It is a form of undirected graphical model in which each edge represent conditional dependencies between random variables at the nodes. Each random variable Y_i is conditioned on an input sequence X. The conditional dependency of the random variable on X is normally represented by some feature functions [4] [5]. This feature function varies according to the application. CRF is commonly used for the labeling of natural language text or biological sequences. They were first used for the task of shallow parsing by Lafterly et al (2001) where CRF were mainly applied for Noun Phrase (NP) chunking. In CRF, with respect to figure 4, Y is dependant only on X while high order CRF or Extended CRF represent a model in which each Y_i is dependants on X as well as on n number of previous variables Y_{i-n}, …., Y_{i-1}

Regular Expression is the standard notation for characterizing text sequences. Finite State Automata (FSA) is a mathematical device used for implementing texts represented by regular expression. A variation of FSA called a Finite State Transducer (FST) is a machine that reads a string and outputs another string. Formally an FST is represented by a 6-tuple [6]. FST's applications are in speech recognition, phrase chunking, POS tagging etc.

Most of the Question Answering systems require answers which are either nouns, adjectives, adverbs or phrases. Deriving these NEs from large collection of documents is a difficult task. Currently there are QA systems available in different languages where they are using keyword extraction techniques.

2. Related Works

NER is a process of finding mentions of specific things in running text. It is an essential tool for QA and IR. But research indicates that NER systems are brittle meaning that NER systems developed for one domain do not typically perform well on another domain. Various approaches available for solving such problems are statistical machine learning techniques, rule based systems and hybrid approaches.

Machine learning methods are using either supervised learning or unsupervised learning techniques. Statistical methods require large amount of manually annotated training data. Few commonly used statistical methods are Hidden Markov Model (HMM), Maximum Entropy Model (MEMM) and Conditional Random Field (CRF). Sequence labeling problem can be solved very efficiently with the help of HMM. The conditional probabilistic characteristics of CRF and MEMM are very useful for the development of NER systems. MEMM is having a POS label biasing problem. But all machine learning techniques require large relevant corpuses which is unavailable in Malayalam. Machine learning methods are cost effective and no need of much language expertise. In [7] authors describe a NER system using CRF. This system uses different contextual information of the words along with both language independent and language dependant features. Paper [8] proposes a HMM based on the mutual information independence assumption where they claimed that their system reaches 'near human' performance. NER system based on MEMM is presented in [9].

Grammar based techniques are used for creating NER systems that obtain better precision but at the cost of lower recall and months of work by experienced computational linguistics. Rule based approaches lack the ability of coping with the problems of robustness and portability. Each new source of text requires significant tweaking of rules to maintain optional performance and the maintenance cost is quite high. Rule based systems performs the best especially for specialized applications. [10] Introduces a rule based system that use handcrafted rules and this approach gave them better performance than the CRF method.

Hybrid Methods either use combinations of different machine learning methods or combinations of rule based and machine learning methods. [11] Presented a tool for the recognition of NE in Portuguese. It has two components-rule based components for recognition of number expressions and hybrid component for names.Lot of work has been reported in the field of NER for English and European languages where one of the main features used is capitalization which is not present in Malayalam Language.

3. Malayalam Named Entity Identifier

NER is the process of identifying and categorizing names in text. In the taxonomy of computational linguistics NER falls under the domain of IE which extracts specific kinds of information from documents as opposed to more general task of document management which extracts all of the information found in a document.

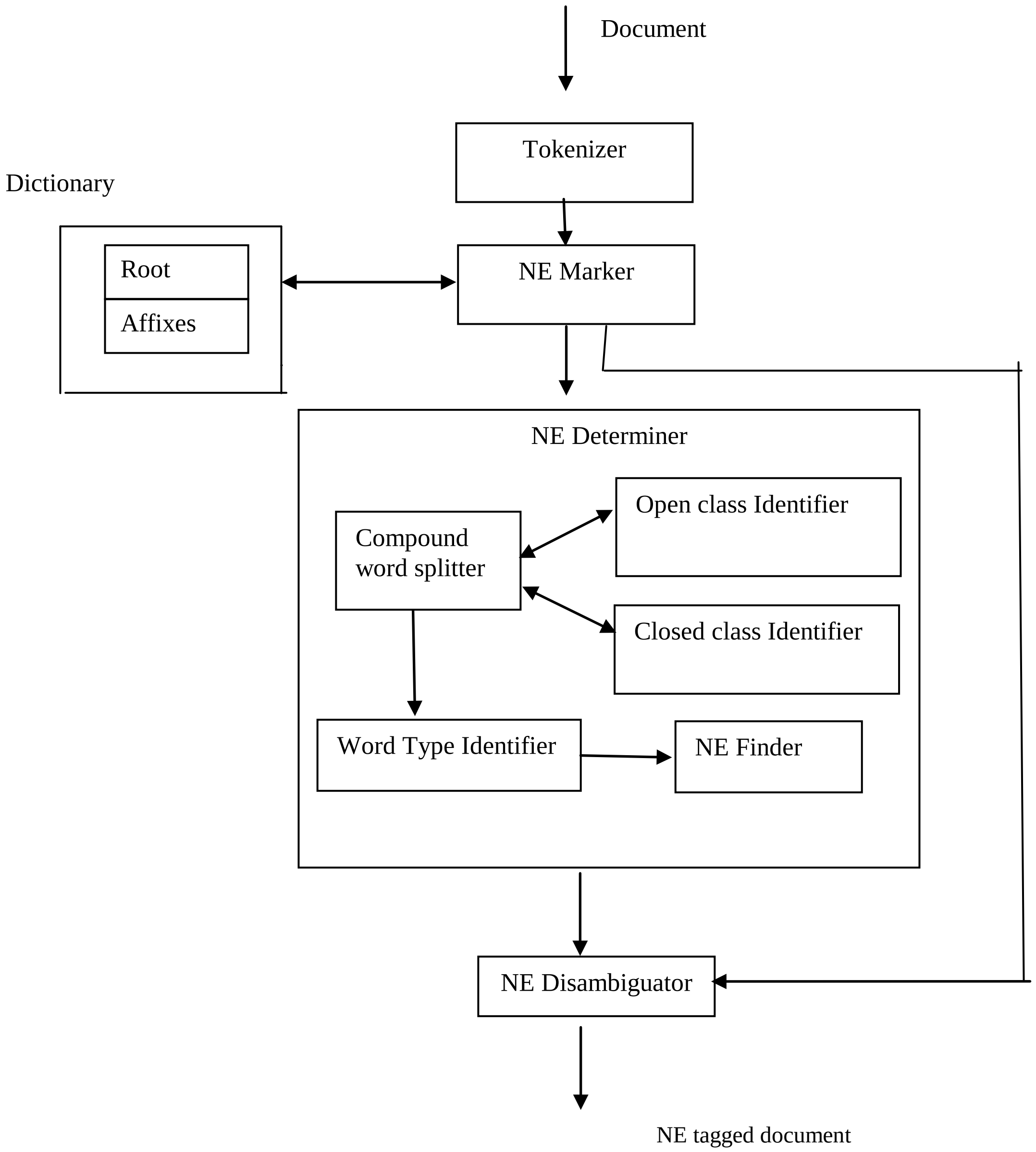


Figure 1: Named Entity Identifier

NE’s are identified by using phonological, morphological, semantic, and syntactic properties of linguistic forms and that act as the targets of linguistic rules and operations. Two kinds of features that have been commonly used are internal and external, internal features are provided from within the sequence of words that constitute the entity-, in contrast, external features are those that can be obtained by the context in which entities appear [12].Based on the above investigation we have categorized an entity as either sole-entity, Constituent-entity, Dependant-entity or Not-an-entity.

3.1 Tokenizer

Input to the Tokenizer block in Fig 1 is a document in Malayalam. During the tokenization process each sentence of the document is taken and split into words or co-occurrence patterns.

3.2 NE Marker

This block checks each token to see whether it is present in the lexicon or not. Lexicon has all the root words along with its POS information. If it is present in the lexicon then it is a simple word, then the word details are retrieved from the lexicon. Based on this information token is marked with the possible NE tags.

Also, a word in the dictionary has information about their possible roles or named entities. But it is not possible to include all proper nouns in the dictionary. And another factor is that 85% of the words in Malayalam texts are compound words. Therefore obtaining NE tags from dictionary practically impossible.

3.3 NE Identifier

If the token M is a compound word then it is to be decomposed into its constituents M_1 to Mi.To find each constituent, the longest match method is adopted. When one component Mi is separated the remaining portion is sent to modification algorithm. The component M is searched in the lexicon, if it is not found transformation algorithm is called to obtain various forms of M and again searching is carried out. If not found process is repeated with next smaller string M. Based on the constituents, NE Identifier assigns suitable tags to each token [13].

Methodology - Finite State Transducer (FST)

Formally, a finite transducer T is a 6-tuple (Q, Σ, Γ, I, F, δ) such that: Q is a finite set, the set of states;
Σ is a finite set, called the input alphabet;
Γ is a finite set, called the output alphabet;
I is a subset of Q, the set of initial states;

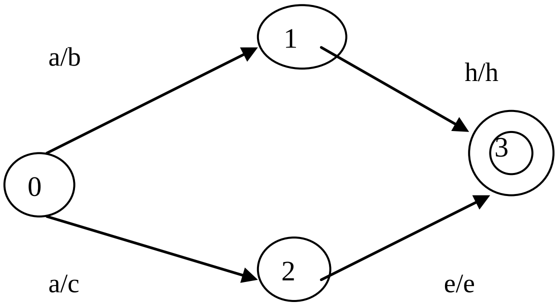


Fig 2 Finite State Transducer

F is a subset of Q, the set of final states; and δ is the transition relation.
Representation of the transducer in Fig.2 is
T=({0,1,2,3},{a,b,c,h,e},0,{3},{0,a,b,1},{0,a,c,2},
{1,h,h,3}, {2,e,e,3})
FST is a machine which accepts a string and translates it into another string. FST can also be used for generating and checking sequences [6]. A compound word is a string of Malayalam characters. To split this string into substrings an FST can be used.
Compound word splitter uses an FST with the following definition.
In the 6-tuple, set of states Q = {A,B,C,D,E,F,G,H}
Initial state I= {A}
Final states F= {C,D,E,F,G,H}
Input alphabet Σ = {compound words}
Output alphabet Γ = { valid simple Malayalam words}
Transition function δ= {NOUN,VERB,ADJECTIVE….,SUFFIX}
FST for compound word splitter is given in FIG.3

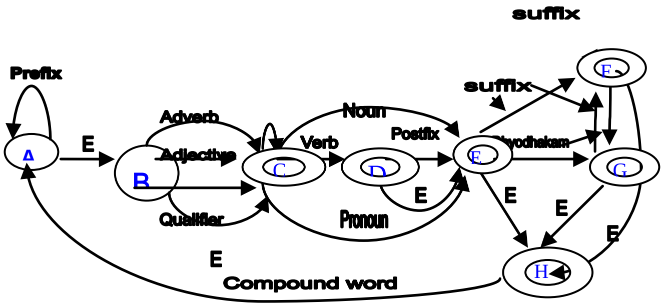


Fig.3 FST for Compound Word Splitter

This system operates in optimal time since the time to assign the tag to sentence corresponds to the time required to follow a single path in a deterministic finite state machine.

3.4 NE Tag Disambiguator

Previous blocks assigns each input token a single/multiple NE tags. Tokens with multiple tags are sent to the Disambiguator to solve the tag ambiguity which removes all tags except one. Output of tag Disambiguator is a string of all tokens along with their NE tags.

Methodology: Extended Conditional Random Field

Tag Disambiguator is implemented using high order CRF or extended CRF. It is an undirected graphical model in which each vertex represents a random variable whose probability distribution is to be inferred and each edge represents a dependency between two random variables. CRF’s avoid the label bias problem, a weakness exhibited by MEMM. The primary advantage of CRF’s over HMM is their conditional nature [4],

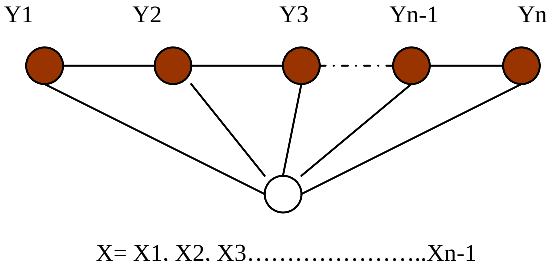


Fig4 Graphical structure of chain-structured CRFs

Let X={X1 …XN} and Y= {Y1…YN} be two sets of random fields. For the given input sequence X, Y represents a hidden state variable and CRF’s define conditional probability distributions P (Y|X) over the input sequence. Sometimes the conditional dependency of each Yi on X will be defined through a fixed set of feature functions (potential functions) of the form f (i, Yi-1, Yi, X). The model assigns each feature a numerical weight and combines them to determine the probability of a certain value for Yi. CRF’s can contain any number of feature functions and the feature function can inspect the entire input sequence X at any point during inference. CRF’s are extended into high order models by making each Yi dependant on a fixed number of previous variables Yi-o ... Yi-1.

NE tagging can be modeled as a sequence labeling task where X= X1X2X3...Xn represents an input sequence of words and Y= Y1Y2Y3...Yn represents corresponding NE Label sequence. The general label sequence Y has the highest probability of occurrence for the word sequence X among all possible label sequences that is Y = argmax{Pr (Y|X)}. These labels are determined by the feature functions.

Main features for NE tagging have been identified based on the word combination and word context. The features also include prefix and suffix for all words [5].
Following are the features used for NE tagging in Malayalam.

- Constituents of current word: These determines the POS tag of the word as noun, verb etc.
- Context word features: Preceding (pw) and following words (nw) of the current word. We have taken pw1, pw2, pw3, nw1, nw2, nw3 as the feature.
- POS information: POS of previous words and in ambiguity resolution, POS of the following words are helpful.
- Contains digits or symbols. If the word contains digits they are marked with ‘NUMBER’ POS (cardinal number(CN) or ordinal number(ON))
- Lexicon feature: It contains Malayalam root words and their basic POS information such as noun, verb, adjective, adverb etc.
- Inflection lists: After analyzing various classes of words inflection lists of nouns, verbs and participles are prepared to improve the performance of the POS tagger.

Working of Named Entity Identifier

1. Tokenize the document.
2. Check each token whether it is present in the dictionary, if it is a simple word then retrieve the word details and determine its NE category
3. If not present in the dictionary and if it is a compound word, call a compound word splitter to obtain its constituents. Use this information to find out the NE type of the current word
4. If above two steps are not sufficient to determine the NE type, call NE detector using ECRF
5. Repeat steps 2-4 for all the tokens

NE Marker determines whether a token is a sole entity or not.NE Determiner marks a token with constituent-entity tag and Tag Disambiguator with Dependant-Entity tag. All other tokens are labeled with Not-an-Entity tag.

4. Tests and Discussions

NER is designed and implemented using J2SDK1.4.2 and MySQL. Its performance is evaluated using standardized

Table 1: Performance of NE Identifier

Token Type	***NE/NAN***		
	Precision	***Recall***	***F-Measure***
Proper Noun	60.0	73.0	65.86
Pronoun	85.2	87.4	86.29
Common Noun	81.4	83.5	82.44
Locative	86.3	85.0	85.64
Accusative	87.6	88.1	87.85
Dative	84.1	87.2	85.62
Instrumental	89.3	88.3	88.79
Reason	81.0	82.4	81.69
Sociative	91.7	89.0	90.33
Car-Num	90.6	91.2	90.89
Ord-Num	93.1	92.3	92.69
Adj-Num	89.0	87.5	88.24
Adj-Quantity	86.0	87.3	86.64
Other Tokens	92.0	90.5	91.24

techniques precision, recall and F-score where Precision is defined as a ratio of number of correct NER tags to the number of NER tags in the output and recall is the ratio of number of correct NER tags to the number of NER tags in the test data. F-score = 2*recall*precision/ (recall+ precision) [14].

Documents related to five different fields are selected as the corpus. Then we randomly selected 8000 sentences for training and 2000 sentences as test set. Precision, recall and F-score obtained for various types of NE tags are shown in table 1. We could overcome the following challenges raised by the Malayalam language features by considering the word level and phrase level information ie by morphological analysis, POS tagging and phrase chunking.

Agglutinative Nature

Malayalam is a highly inflectional and agglutinative language. 85% of words in Malayalam text are compound words and hence role of these words can be decided only by knowing its components and their types. Role of an entity depends on the importance of the word which is decided by local and global information. To derive local information, each word is analyzed and collected its component details.

Word Order

Malayalam sentence is a sequence of words where words may appear in any order and each word can be a combination of any number of stems and affixes. Even though there is no specific order for the words in the sentence, within a chunk word categories are related.

In Malayalam language there is no distinction between uppercase and lowercase. Hence proper techniques are to be adopted to overcome such challenges.

5. Conclusion

Main task of NER is to identify and classify named entities in a given document. Identification is concerned with marking the presence of a word/phrase as NE in the given sentence and classification is for denoting the role of the identified NE.

Most of the systems have concentrated on three kinds of NEs ie on the roles of proper nouns, Time and percentage expressions. But these entity types are not sufficient for many question answering systems where entities like reason, cause, instrument etc are to be identified. The NER system described here is designed incorporating these types. Also this paper addresses the problem of NER in a query which involves the detection and classification of the named entity in a given query into predefined classes.

We have selected formal text since this is developed as a part of QA system based on health IR. For this application text from various textbooks, journals and magazines and web sites are selected which are mostly formal texts.

6. References

[1] Stefan Schwarzler Joachim Schenk,Frank Wallhoff and Gunther Ruske,"Natural Language Understanding by Combining Statistical methods and Extended Control Free Grammars",Proceedings of 30th DAGM Symposium on Pattern Recognition,Springer-Verlag Berlin,Heidelberg,2008.

[2] Lev Ratinov Dan Roth, "Design Challenges and Misconceptions in Named Entity Recognition", Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL), pages 147–155, Boulder, Colorado, June 2009

[3] A .R .Rajarajavarma,"Keralapanineeyam", National Book Stall, Kottayam, 2000.

[4] Hanna.M.Wallach, "Conditional Random Fields", University of Pennsylvania CIS Technical Report MS-CIS-04-21.

[5] Samir AbdelRahman, Mohamed Elarnaoty, Marwa Magdy and Aly Fahmy," Integrated Machine Learning Techniques for Arabic Named Entity Recognition", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 4, No 3, July 2010

[6] Bindu.M.S, Sumam Mary Idicula,"Analysis of Malayalam compound words and Implementation of a compound word splitter tool using Finite State Models", International Conference on Modeling and Simulation India 1-3 December 2009.

[7] GuoDong Zhou Jian Su ,” Named Entity Recognition using an HMM-based Chunk Tagger”, “Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 473-480.

[8] Mohammad Hasanuzzaman1, Asif Ekbal2 and Sivaji Bandyopadhyay3,” Maximum Entropy Approach for Named Entity Recognition in Bengali and Hindi”, International Journal of Recent Trends in Engineering, Vol. 1,No.1, May 2009

[9] Burr Settles,” Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets”, Proceedings of the COLING 2004 International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA). Geneva, Switzerland. 2004.

[10] Gjorgji Madzarov, Dejan Gjorgjevikj and Ivan Chorbev,” A Multi-class SVM Classifier Utilizing Binary Decision Tree”, Informatica 33 (2009) 233-241

[11] B. Sasidhar, P. M. Yohan,Dr. A. Vinaya Babu, Dr. A. Govardhan,” Named Entity Recognition in Telugu Language using Language Dependent Features and Rule based Approach”, International Journal of Computer Applications (0975 – 8887) Volume 22– No.8, May 2011

[12] Xiaofeng Yu,”Chinese Named Entity Recognition with Cascaded Hybrid model”,Proceedings of NAACL HLT 2007 Companion Volume, pp 197-200,April 2007

[13] Asif Ekbal and Sivaji Bandyopadhyay, “Named Entity Recognition Using Appropriate Unlabeled Data, Post-processing and Voting “,Informatica 34 (2010) 55–76

[14] Alireza Mansouri, Lilly Suriani Affendey, Ali Mamat, “Named Entity Recognition Approaches”, IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.2, February 2008

Bindu.M.S received her B.Tech degree from M.A College of Engineering, Kothamangalam in 1986 and M.E degree from Coimbatore Institute of Technology, Coimbatore in 1988.She is currently pursuing the Ph. D. degree in the research area of Natural Language Processing from Cochin University of Science and Technology, Cochin, India.

During 1988-1998 she was with Manipal Institute of Technology, Manipal, as Lecturer and then as Reader in the Department of Computer Science and Engineering. Currently she is working as Reader in the Department of Computer Applications with Mahatma Gandhi University, Kottayam India. She has published several papers in International and National conference proceedings. Her research interests include Natural Language Processing, Artificial Intelligence and Information Retrieval.

Dr. Sumam Mary Idicula took B.Sc (Engg) degree in Electrical Engineering from College of Engineering Trivandrum in 1983. She pursued her Master studies in the field of Computer and Information Science in Cochin University of Science & Technology and took M.Tech degree in 1986. She started her carrier as lecturer in the Department of Computer Science of Cochin University of Science & Technology in 1987. She took PhD degree in Computer Science later and is now working as Reader in the same Department.

She is an active researcher in the field of Natural Language Processing and Human Computer Interaction. She has undertaken 3 major projects supported by ISRO and UGC in the field of Natural Language Processing and 2 major projects supported by AICTE and KSCSTE in the field of Human Computer Interaction. She is guiding several M.Tech students & Ph.D Scholars. About 40 research papers have been published by her in the field of Computer Science in reputed journals and in international conferences. She has visited Europe and United States for participating in International Conferences & Workshops.

She is a member of the Board of Studies of Computer Science and Board of Studies of Computer Applications of Cochin University of Science & Technology and also a member of the Academic Committee of CUSAT.

Efficient Online Tutoring Using Web Services

Mr.M.Balakrishnan[1] and Dr.K.Duraiswamy[2]

[1] **Assistant Professor, Department of Computer Science and Engineering,
Selvam College of Technology, Namakkal, Tamilnadu, India**

[2] **Dean (Academic), K.S.Rangasamy College of Technology,
Tiruchengode, Tamilnadu, India**

Abstract

Web Services is a technology that allows applications to communicate with each other in a platform and programming language-independent manner. The development and implementation of a tutoring service through online is presented. Online tutoring is all about using web services or the internet for tutorials or tutoring activities. Students would be learning from their tutors through the use of internet. Online tutoring or e-tutoring would need certain applications and programs, like instant messaging that would make discussions possible. Aside from discussions and lectures done through internet and web conferencing, quizzes, exam results, and recommendations are also done through the internet. In this paper, we propose a structural frame of integrated web services for online interactive training/education environments with object-oriented interface to multiple platforms of resource allocation. After analyzing the requirements tutor environment, it is shown how Web Services can address many of the requirements put forward in terms of operability, deployment and usage services. The proposed framework of is experimented to show the performance improvement with the existing model in terms of application throughput, delivery report, interaction of more number of people, and scalability.

Keywords: *Web service, online Tutoring, Throughput, and Scalability.*

1. Introduction

In using technology to provide tutoring online, new studies are suggesting that the important element may be the definition of the process of tutoring in the new cyberspace environment more than the choice of the technology. Although electronic tools are needed to deliver tutoring online, a definition of the online process and its best practices may be needed first in order to help select the appropriate technology or, in the words of Frank Christ (2002), put “pedagogy before technology.” Historically, online tutoring began with email. In this format, a student sent a question to the tutor with the expectation that the return email would contain the answer. Instead, what happened was a disconnection: The tutor, being a good guide, sent back a Socratic answer with more questioning prompts; the student, expecting the answer, became frustrated. Although the student may expect a give and take interaction in a face-to-face tutoring session, the email format suggested to the student that the question should be answered with a direct answer. This illustration is an example of using technology without fully developing the concept of tutoring in the online environment.

Initial tutoring online models begin with email but there has been an emergence of new models as new tools became available, both in asynchronous and synchronous formats. The following models are presented in the context of the specific tool used, for example, Blackboard or NetTutor.A learning service is in our understanding an event. In order to support the accomplishment of a specific educational objective event is provided by a learning service provider. This is achieved by creating a learning environment consisting of educators, educational material, communication infrastructure, meeting places, etc. Examples of a learning service are the delivery of a course, the provision of a web-based training application or the provision of self-study material.

We envision a scenario where learning services are announced and mediated by electronic means. Although the web enables choosing from a variety of educational resources, it is difficult for learners to find appropriate learning services such as courses, seminars, and web-based training applications. Corporate and independent learners seek learning services with heterogeneous properties (traditional courses, online courses, assessment services, mini-learning units, etc.) from heterogeneous sources (in-house training, external training providers, higher education institutions, etc.). The rationality of the selection process of a learning service performed by a human being is limited for the following reasons, limited overview of the learning services available, limited capabilities of processing all the information describing learning services.

2. Literature Survey

In people’s daily life, with the development of human-computing interaction, more and more natural human-computing interfaces have been integrated for enhancing work efficiency [1]. In the learning area, human-computing interface can facilitate the teacher to teach the class and help the students learn and discuss with others. Several projects, such as [2] and [3], implement special human-computing interfaces in their learning environment. At the same time, mobile devices such as mobile phone, smart phone, PDA, and laptop have been easily accessible for ordinary people. Researchers in [9] and [11] emphasize that mobile devices play an important role in learning. For example, the teacher uses his Smart Phone to bring the presentation file and to control the slideshow, while the students can use a laptop to discuss with others. Some of these features have been incorporated in several projects.

Previous Smart Platform enables mobile devices roaming with users to connect into Smart Space by preinstalled modules (eContainer and eADK-based agent of Smart Platform). However, it lacks convenience for the users, especially for those who first come into Smart Space to use their mobile devices. Open Smart Platform applies OSPG as the Web-based mobile interface for mobile devices interaction in Smart Space. OSPG provides the mobile interfaces, such as PPT upload or Turn-to-Next-Page, as a Web page [5], [7].

Since almost all the mobile devices such as laptop, PDA, smart phone, or even normal cell phones have an integrated Web browser, it is very convenient for the users to access the services and interfaces inside Smart Space [4], [8]. Besides providing extensibility, load scalability for the mobile devices to interact with Smart Space is also improved by this mechanism. Serving as a proxy between the mobile devices and modules inside Smart Space, when the concurrent mobile device access increases, OSPG could involve load balancing [6], [10] and cache mechanism to alleviate the total load for Smart Space and also could control the total number of the concurrent mobile devices in order to avoid the overload of the whole system. This article will review the evolution of online tutoring and discuss the best practices suggested by the studies.

3. Efficient Online Tutoring Service

Tutoring services are complete entities designed for a specific purpose and targeted at a specific audience. Providers of learning services can state clearly which kind of skills they want to develop and train in the learner. Learning objects are of a more general nature and of a smaller granularity level. Educators and (semi-) automated tutoring systems compose learning services out of learning objects and other educational resources. Because of the extensive use of resources, learning services - especially in the corporate world - do not come for free. Hence, exchange transactions comprising provision, offer placement, announcement, booking, and payment of educational services need to be supported by a mediating infrastructure where users are authenticated.

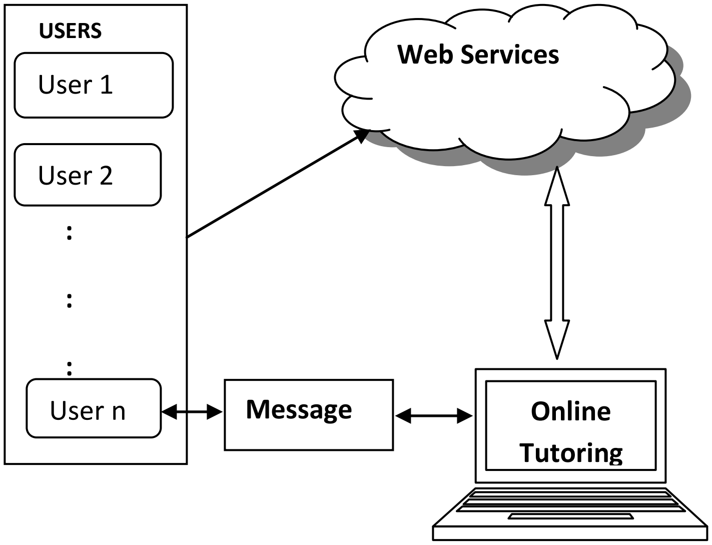


Fig 1. Online Tutoring Service

Tutoring services which make use of physical or human resources are offered according to a specific schedule since the use of those resources needs to be managed. A talk is held at a specific point of time, a course is offered within a semester period, tutoring sessions require an appointment, etc. When it comes to the delivery of a learning service, providers follow a specific objective. In the case of the delivery of a course, for example, the accompanying objective can be explicitly expressed by the educational objective and the learning goals of the course. Consumers of services are motivated by a particular objective when they consume a service. Mediation of learning services requires matching the goals of the prospective learner with the educational objective addressed by a learning service. The modeling of learning services with web services also opens the possibility of automated integration of educational services into a smart learning space and the automated combination of them. To perform this task, however, we need also semantic information about the educational services.

The online tutoring service was originally conceived as a means to provide assistance for students enrolled in distance learning classes. However, before its launch, the program was opened up as a free resource available to all Pima students. Tutoring is provided in writing, math, and accounting. Students who wish to use the service obtain an access code from their instructor or the West Campus Learning Center. Students may choose to communicate with a tutor in real time or may leave questions and retrieve the tutor’s response at a later time. The program provides a number of options: one-on-one tutoring; group sessions led by a tutor; and sessions that include groups of students, their instructor, and a tutor for support. As online tutoring continues to grow, there is an expectation that more academic subjects will be added to the service. After Pima's yearlong pilot project, the tutors are now developing best practices, which focus not on the technology itself, but on the development of the whole online tutoring environment. Moreover, part of establishing the online tutoring environment is to establish the expectations and the parameters for students.

4. Result and Discussion

Web Services provide an ideal contemporary solution for hyper linking software components over the Internet. During the 11 weeks of the course, a total of 25 users participated in the experimental group. In experiment initially 5 users only participated. The communication rate was 82%.First of all, the student judgments should be in line with the staff-tutors' rating. After all, the students could have been satisfied too easily. The two staff-tutors rated all questions, including the questions started but not yet rated by the students. The overall agreement between the tutors on solved versus not-solved questions is high: 85% or 73%.

If we combine the judgment of the students and the tutors, by counting a question as solved if at least two of the three ratings are 4 or above, the number of questions solved is approximately the same as the number indicated by the students. So student opinion does not differ much from expert (staff) opinions. Second, irrespective of an overall agreement between students and staff, there should only be very few 'false-positives'. A false-positive is an answer that according to the student is right but actually is wrong. Too many false-positives are a threat to the quality of education. Based on the ratings, we identified 8 questions that required further analysis.

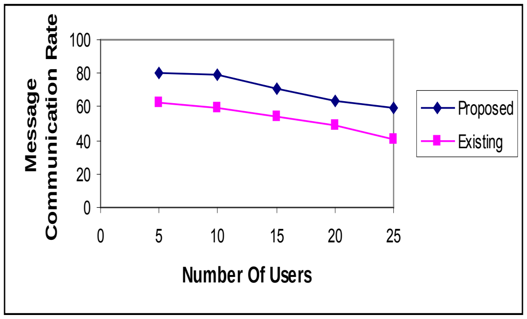


Fig 2. Number of users Vs Message Communication rate

From the experimental results, if Number of users increased the communication rate decreased. Compared to existing work our proposed model is having high efficiency.

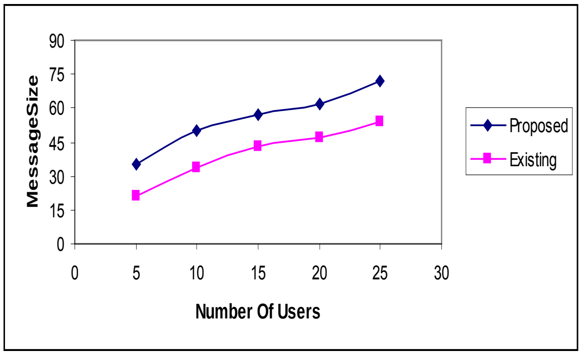
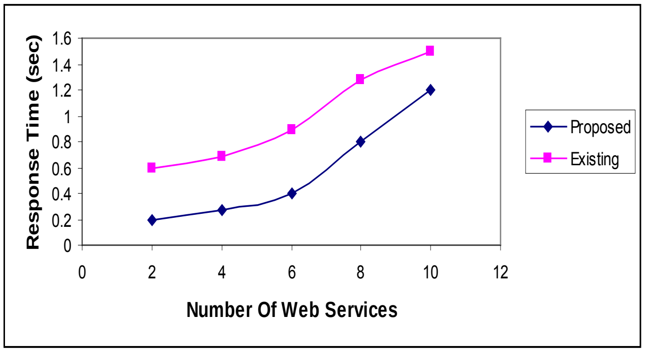


Fig 3. Number of users Vs Message Size

The above figure shows the figure as Number of users Vs Message Communication rate. The number of users is directly proportional to the message size. So if more number of users participates the message size also increased automatically.

Fig 4. Number of web services Vs Response Time



The figure 4 shows that the number of web services vs. response time. If Number of web services increased, then the response time also increased. Our proposed model is having effective response time compared to existing model.The proposed framework is experimented to show the performance improvement with the existing model in terms of application throughput, delivery report, interaction of more number of people, and scalability. Our framework has high throughput, less delivery time and increased scalability. Our model interacts with more number of users.

5. Conclusion

The evolution of online tutoring has showed us that success may not depend so much upon the tool selected, but on the development of an appropriate culture for online tutoring, an understanding of the process and parameters involved. Learning services which make use of physical or human resources are offered according to a specific schedule since the use of those resources needs to be managed.

In this paper we propose an encapsulated education environment which effectively integrated Web Services to exploit the resource sharing for relevant online tutoring domain. Our model has some properties as Easy installation, Ease of use, Low maintenance efforts, and Integration with other Internet/Intranet based education tools. The experimental results show the improved performance of throughput, delivery time and scalability.

References

[1] Yue Suo, Naoki Miyata, Hiroki Morikawa, Toru Ishida, and Yuanchun Shi, “Open Smart Classroom: Extensible and Scalable Learning System in Smart Space Using Web Service Technology”, IEEE transactions on knowledge and data engineering, vol. 21, no. 6, june 2009.

[2] H. Allert, C. Richter, and W. Nejdl. Learning objects and the semantic web. explicitly modelling instructional theories and paradigms. In Proceedings of E-Learn 2002: World Conference on E-Learning in Corporate, Government, Healthcare, & Higher Education (formerly the WebNet Conference), Montreal, Canada, Oct. 2002.

[3] ebXML, Business Process Specification Schema.

http://www.ebxml.org/specs/ebBPSS.pdf.

[4] N. Friesen, A. Roberts, and S. Fisher. Cancore: Metadata for learning objects. Canadian Journal of Learning and Technology, 28(3):43–53, 2002.

[5] A. Maedche and S. Staab. Services on the move – towards p2p-enabled semantic web services. In Proceedings of the Tenth International Conference on Information Technology and Travel & Tourism, ENTER 2003, Helsinki, Jan. 2003.

[6] M. Sintek and S. Decker. Triple–a query, inference, and transformation language for the semantic web. In Proceedings of the International Semantic Web Conference (ISWC2002), Sardinia, Italia, June 2002.

[7] Addison Wesley. (2005). The Tutor Center. Retrieved September 20, 2005 from http://www.aw-bc.com/tutorcenter/index.html

[8] Christ, F.L. (2002, February). Achieving student retention, satisfaction, and success through online pedagogy. A presentation at TechEd, California State

University, Long Beach.

[9] Doherty, B. & Atkinson, M. (2004, Spring). A pilot study of online tutoring using Smarthinking. PowerPoint presentation. Retrieved June 15, 2004 from http://www.smarthinking.com

[10] Fryer,W. (2003, June 30). John Couch: Delivering measurable achievement, NECC Presentation. Retrieved April 24, 2006, from http://webpages.acs. ttu.edu/wfryer/ necc2003/couch.html

[11] L. Stojanovic, S. Staab, and R. Studer. elearning based on the semantic web. In Proceedings of the World Conference on the WWW and Internet WebNet2001, Orlando, Florida, USA, 2001.

M.Balakrishnan received the M.E. degrees in Computer Science and Engineering from K.S.Rangasamy College of Technology, Tiruchengode, in 2006 respectively. During 2007-2009, he worked as Lecturer in K.S.Rangasamy College of Technology in Tiruchengode. He now with Selvam College of Technology, Namakkal, Tamilnadu, India as Assistant Professor in Department of Computer Science and Engineering.

Dr.K.Duraiswamy received the B.E., M.Sc. and Ph.D. degrees, from the University of Madras and Anna University in 1965, 1968 and 1987 respectively. He worked as a Lecturer in the Department of Electrical Engineering in Government College of Engineering, Salem from 1968, as an Assistant professor in Government College of Technology, Coimbatore from 1983 and as the Principal at K.S.Rangasamy College of Technology from 1995. He is currently working as a Dean in the Department of Computer Science and Engineering at K.S.Rangasamy College of Technology (Autonomous Institution).His research interest includes Mobile Computing, Soft Computing, Computer Architecture and Data Mining. He is a senior member of ISTE, IEEE and CSI.

Improved Datagram Transport Protocol over Wireless Sensor Networks- TCP Fairness

Senthil Kumaran .M [1] and Dr. R. Rangarajan [2]

[1] Professor/CSE
Muthayammal Engineering College
Rasipuram, India.

[2] Dean/ECE
Dr. Mahalingam College of Engineering & Technology
Pollachi, India.

Abstract

TCP connections have small bandwidth-delay product and frequent packet loss in wireless sensor networks due to route breakages and radio interference. Datagram transport protocol provides a reliable end-to-end transport protocol over wireless sensor networks. This paper deals with improvement of TCP Fairness as Fairness in wireless sensor networks plays a major role to have maximum fair share of available bandwidth among the nodes, thus energy is consumed. A distributed adaptive max-min algorithm has been proposed in order to improve the fairness in WSNs. The proposed scheme incorporates two techniques: a fixed-size window-based flow-control algorithm and a cumulative bit-vector-based selective ACK strategy. Security has got the major impact over WSNs and that has been overcome by logical Tunneling. The simulation results show the improvement in terms of fairness, throughput and delay and packet loss using Network Simulator NS-2.

Keywords: *Fairness, Congestion Control, Bandwidth Delay Product, Dynamic Source Routing*

1. Introduction

Recent advances in wireless communication technology and portable devices have generated much interest in wireless sensor networks (WSNs). A WSN is a collection of wireless devices moving in seemingly random directions and communicating with one another without the aid of an established infrastructure. It is characterized by low variable bandwidths, high variable delays, significant non congestion-related packet losses, and occasional communication blackouts due to route breakages. Under such operating conditions, the traditional reliable transport-layer protocol TCP, which is used widely in the wired network, is not suitable for WSNs. The proposal of this work is carried out in the direction of analyzing congestion control algorithm and the strategy used to guarantee reliable delivery. Bandwidth-delay products (BDPs) represent the maximum amount of unacknowledged data that are allowed in flight at any moment in the network. When the amount of traffic injected by a source exceeds the BDP of the connection, the excessive packets that are queued in the network lead to undesirable queuing delays and congestions. Correspondingly, the network cannot be fully utilized when the total packets in the network are less than the BDP. As such, the BDP plays an important role in the design of a congestion control algorithm for the effective usage of network resources. For a window-based transport protocol like TCP, the transmission window size must be carefully adjusted for networks with different BDP values. TCP was originally designed for a general wired network where the BDP is not very large and packet losses rarely occur. However, with the emergence of various types of networks such as high-speed and satellite networks, TCP can no longer guarantee good bandwidth utilization. It is desirable that the system knows the BDP for each connection in advance so that it can limit the amount of traffic pumped into the network and maintain the optimum throughput.

In traditional wired networks, the TCP source is unaware of the available BDP and dynamically determines it by creating congestion during the transmission. However, in WSNs, the BDP of a path can be determined in advance by using routing protocols such as DSR (Dynamic Source Routing Protocol) and AODV (Adhoc On Demand Vector Routing Protocol). This allows the transport layer to intelligently set its transmission window before connection establishment. In this improved DTPA, the transmission window is fixed at a small value, which is as low as several packets. This proposed mathematical model calculates the optimum transmission window size, and its value has been found to be equal to the

BDP of the path plus 3. More effort should be put into the ability of the scheme to quickly detect and recover the lost packets.

The currently prevalent Reno variant of TCP employs a cumulative ACK scheme, through which packet losses are detected with at least three duplicate ACKs, i.e., four identical ACKs without any other intervening packets. Using this technique, networks with high packet losses cause TCP to rely heavily on time-outs to detect packet losses, which degrades the TCP performance significantly. The deficiency of TCP Reno's ACK scheme has led to the design of the selective ACK (SACK) scheme, through which the ACK packet can carry both the negative and positive information of packet transmissions. In addition, the TCP option may be used to carry other information such as time stamp. As such, the number of usable SACK blocks is further reduced, and the TCP SACK scheme is unsuitable for use in networks with frequent packet losses. In the proposed scheme, cumulative bit-vector based SACK scheme is deployed, through which each bit in the vector represents the reception status of one packet. The bit value and the bit position in the vector are used to predict packet losses. Hence, each ACK packet can acknowledge a wide range of packets by using small overheads. Correspondingly, this transport protocol essentially becomes a datagram-oriented protocol.

2. Literature Review

Various solutions have been proposed to provide reliable packet delivery in WSNs in recent years. Most of the existing proposals in WSNs focus on the modification of congestion control algorithms in the transport protocol. These proposals can be divided into two categories i.e., non-TCP variants and TCP variants. The schemes in [5] [6] and [7] belong to non-TCP variants, where the source adjusts the transmission rate based on explicit feedback from intermediate nodes along the path. Although relatively accurate congestion information can be obtained, the algorithms in these schemes cannot retain the end-to-end semantics of a transport protocol. Moreover, they require complicated mathematical computation and incur excessive network overhead.

In contrast with non-TCP variants, the majority of the proposals are modified versions of the legacy TCP protocol. In [8] and [9], the strategies proactively detect incipient congestion by relying on some measured metrics at the source, such as the variation of the measured round-trip times (RTTs) and short-term throughput, since packet loss information alone cannot provide an accurate congestion indication. In [10] [11] and [12], the authors attempt to minimize the contentions between data and ACK packets by adaptively reducing the number of ACK packet transmissions in the network. The drawback of these TCP variants is that they still adopt the AIMD congestion control algorithm with unnecessary large transmission windows. Chen et al. [13] propose a scheme where the TCP sender limits the size of its maximum transmission window to the BDP of the path in multi-hop networks. However, the small transmission window leads to high AIMD costs, and the TCP source is also unable to detect the packet loss

3. Improved DTPA-TCP Fairness Scheme

The improved DTPA works on the basic principle of DTPA that utilizes selective acknowledgement scheme and in addition uses max-min fairness algorithm to improve the TCP fairness criteria.

A. Strategies in Flow Control Scheme

In this paper the congestion control is done in the WSNS uses datagram based technique, fixed window flow control and cumulative bit vector SACK strategy.

1) Datagram-Based Technique:

In case of datagram based technique, each datagram is sequenced. The sequence number in each data header does not represent the highest byte of that data like in TCP. With a datagram protocol, the IP fragmentation of a packet is highly undesirable during the transmission, because the fragmentation cause inefficient resource usage due to the incurrence of a higher MAC overhead. The loss of a single fragment requires the source to retransmit all of the fragments in the original datagram, even if most of the fragments are received correctly at the destination. Datagram is small enough to avoid fragmentation.

2) Cumulative Bit Vector based SACK:

ACK header illustrates the new ACK strategy. H is the highest sequence number of the datagram that has been received. There is a bit in the header named L to indicate the existence of an out-of-order data. The L flag is turned on whenever an out-of-order segment arrives, which implies that there may be missing packets. The vector field consists of k bits representing the receiving status of a set of earlier packets. Let ai be a single bit that indicates the arrival status of the packet with sequence number H_i. Use the L flag for the realization of a cumulative acknowledgment mechanism. If the L flag in a received ACK packet is turned off, this means that the transmission is in a normal status, without data packet losses or out-of-order transmissions. Hence, an acknowledgment of sequence number H indicates that all datagram up to H have been received.

3) Fixed Window Flow Control:

Improved DTPA employs a sliding window transmission technique. Given the number of hops n, the window size is fixed at

$$w\ (n) = BDP\ (n) + \alpha\ (n)$$

where α(n) is a small value used to guarantee that there are enough packet transmissions and ACK arrivals at the source in case of packet losses. For instance, with w = BDP = 1, the source cannot detect the packet loss through the ACK, because there is only one packet transmission, and it is the one that is missing. A mathematical model is developed to calculate the throughput of a single DTPA flow over an n -hop ad hoc network. The throughput is derived as a function of packet loss rate, path length n, and transmission window size w (n). Determine an appropriate transmission window with which the throughput of an n -hop chain can be maximized.

DTPA relies on the path length information provided in routing protocols such as DSR and AODV to determine the BDP of a path. In DSR, the header of each IP packet carries the instantaneous hop count information. In AODV, each route table entry at the source also contains the hop length. The hop count information provided by these routing protocols can be passed on to the transport layer so that DTPA can intelligently set its transmission window. Find the BDP of a path so that DTPA can be compatible with more routing protocols. The strategy guarantees that the network pipeline is at least fully utilized and that there is no heavy congestion and contention in the networks.

4) Retransmission Mode:

Derive a mechanism for deciding on whether to retransmit the lost packet. DTPA source decides to retransmit a packet by either the receipt of ACKs or a time-out event. The source DTPA depends on the L flag and the bit vector in the ACK header to detect the possible packet losses. It keeps a retransmission buffer for storing the incoming ACK information with a turned-on L flag on a per-connection basis**.** For a time-out event, the DTPA source assumes that there is no outstanding packet in the network. Since the transmission window is fixed at a value greater than one, besides retransmitting those lost packets recorded by the retransmission buffer. The source also transmits new packets if the transmission window allows. In DTPA, the Retransmission Time-Out (RTO) timer works in the same way as that in TCP. Except that it does not exponentially increase itself in the event of retransmission, because the time-out event here implies a window's worth of packet losses rather than a heavy congestion in the system.

5) Performance Metrics for Improved DTPA:

Throughput:

Similar to TCP, the behavior of the DTPA protocol is regarded as a cyclic evolution. It defines one cycle as the interval between the end of one time-out event and the end of the next time-out event. The cycles form a renewal process due to the independent packet losses. The cycle duration is a random variable that is further divided into two non overlapping components at the time when the time-out occurs. It define the DTPA throughput for an n-hop static linear chain

Fairness:

Significant TCP unfairness in ad hoc networks has been revealed. TCP unfairness is mainly attributed to the unfairness of the MAC protocol, which results from the nature of shared wireless medium and location dependency. To solve the TCP unfairness in wireless sensor networks, adaptive max-min fairness algorithm is used in our DTPA model. Fairness criteria for DTPA flows in wireless sensor networks are compared with improved DTPA. In multi-level scheduling for wireless ad-hoc networks the max-min fair allocation of the fair shares is made at the lower-most layer (MAC layer). It mainly lays down the framework to calculate the fair shares that would achieve max-min fairness in an ad-hoc network. Then design distributed algorithms that allow each node to determine its max-min per-link fair share in a global ad-hoc network without knowledge of the global topology of the network.

4. Experimental Performance of Improved DTPA

The simulation of the improved DTPA model is carried out in NS-2 simulator to validate the analytical model and evaluate the performance of DTPA with improved DTPA. The simulations using static topologies interoperate with other WSN DSR and AODV to handle the mobility issues. In the simulation scenarios, all nodes communicate with identical half-duplex wireless radios with a bandwidth of 1 Mbps. The radio propagation model uses free-space attenuation at near distances and an approximation to two-ray ground at a far distance by assuming specular reflection off a flat ground plane. DSR is used as the routing protocol for improved DTPA. The packet size is set to 512 bytes, and fragmentation does not take place during transmission.

To validate the analytical model, we run a DTPA flow over the n-hop static linear chain defined in the model. Each simulation is run for 500 seconds. The data and ACK packets are discarded with probability by the receiving node rather than being dropped. Fig. 1 shows the throughput comparisons between the AODV and DSR-DTPA (improved) simulation and results of the proposed improved DTPA model for different number of nodes. With Fig1 the throughput of the DSR-DTPA model increase when compared to that of the AODV model of transmission.

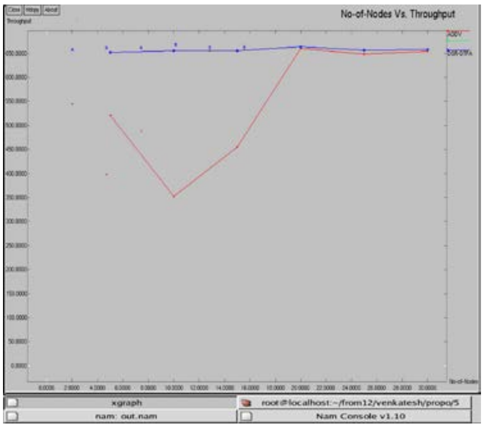


Fig. 1 Comparison of AODV with DSR-DTPA (improved) Number of nodes Vs throughput

The node variation is plotted against the packet loss rate of each transport layer packet as shown in Fig 2. Both results show that the delay decreases with even in the nodes being increased and also noted that delay for DSR-DTPA is less compared to that of the AODV routing protocol. The results from the analysis of the proposed model match closely with the simulations. It can be seen that the delay do not have a significant impact on the number of nodes being increased (Fairness criteria) with throughput performance in multihop wireless sensor networks, as the improved DTPA throughput always increases with number of node variations.

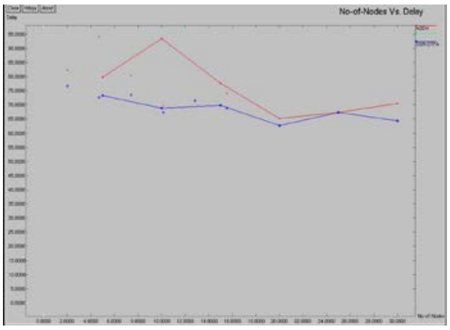


Fig. 2 Comparison of AODV with DSR-DTPA (improved) Number of nodes Vs Delay

Comparing to AODV, the throughput, average RTT, average maximum IP queue size and the number of retransmissions can, respectively, be improved by up to 35 percent. Using a small transmission window, the CWL source sends a limited amount of packets into the networks so that it cannot detect packet losses via the reception of enough ACK packets. The source has to heavily rely on time-out events to detect and retransmit the lost packets, and correspondingly, the CWL goes back to a slow-start phase, with its transmission window dropped to one, which results in throughput degradation. As such, improved DTPA scheme can be utilized in a general ad hoc network for efficient fairness in transmission.

5. Conclusion and Future Work

The improved DTPA model proposed in this work had a reliable Datagram Transport Protocol over wireless sensor networks with better fairness in multi-hop increased node transmission. As the BDP of a path in WSNs is very small, any AIMD-style congestion control algorithm is costly and is hence not necessary for wireless sensor networks. On the other hand, a strategy for guaranteeing reliable transmissions and recovering frequent packet losses plays a more critical role in the design of a transport protocol. With this basis, our scheme incorporates a fixed-window-based flow control and a bit-vector-based SACK strategy with which the ACK packets contain a vector of bits representing the reception status of the set of packets that were transmitted earlier.

With NS-2 simulator, improved DTPA model is evaluated and shown that results improves the network throughput, average RTT, and decreased average delay in the network, and number of retransmissions by up to 35 percent as compared to the AODV. Since DTPA employs a window-based congestion control coupled with an ACK technique similar to TCP, it is believed that DTPA also experiences severe unfairness among competing flows in wireless sensor networks. Hence, in order to provide fairness for DTPA flows in wireless sensor networks, max-min algorithm is deployed in improved DTPA. This solution can be extended to evaluate fairness with large number of nodes in WSNs.

REFERENCES

[1]. Xia Li, Peng-Yong Kong and Kee-Chaing Chua, “DTPA: A Reliable Datagram Transport Protocol over Ad hoc Networks”, *in IEEE Transactions on Mobile Computing*, October 2008.

[2]. M. Mathis, J. Mahdavi, S. Floyd, and A. Romano, “TCP Selective Acknowledgement Options”, *IETF RFC 2018*, 1996.

[3]. R. Kettimuthu and W. Allcock, “Improved Selective Acknowledgment Scheme for TCP,” *in Proceedings International Conference on Internet Computing (IC ’04)*, pp. 913-919, 2004.

[4]. H.-S. Wilson So, Y. Xia, and J. Walrand, “A Robust Acknowledgement Scheme for Unreliable Flows,” *in Proceedings IEEE INFOCOM ’02*, vol. 3, pp. 1500-1509, 2002.

[5]. G. Anastasi, E. Ancillotti, M. Conti, and A. Passarella, “TPA: A Transport Protocol for Ad Hoc Networks,” *Proc. 10th IEEE Symp. Computers and Comm. (ISCC ’05)*, pp. 51-56, 2005.

[6]. K. Chen, K. Nahrstedt, and N. Vaidya, “The Utility of Explicit Rate-Based Flow Control in Mobile Ad Hoc Networks,” *in Proceedings IEEE Wireless Comm. and Networking Conference (WCNC ’04)*, no. 1, pp. 1904-1909, 2004.

[7]. H. Zhai, X. Chen, and Y. Fang, “Rate-Based Transport Control for Mobile Ad Hoc Networks,” *in Proceedings IEEE Wireless Communications and Networking Conference (WCNC ’05)*, pp. 2264-2269, 2005.

[8]. K. Sundaresan, V. Anantharaman, H.-Y. Hsieh, and R. Sivakumar, “ATP: A Reliable Transport Protocol for Ad Hoc Networks,” *IEEE Transactions on Mobile Computing*, vol. 4, no. 6, pp. 588-603, November 2005.

[9]. Z. Fu, B. Greenstein, X. Meng, and S. Lu, “Design and Implementation of a TCP-Friendly Transport Protocol for Ad Hoc Networks,” *in Proceedings 10th IEEE International Conference on Network Protocols (ICNP ’02)*, pp. 216-225, 2002.

[10]. S. EIRakabawy, A. Klemm, and C. Lindemann, “TCP with Adaptive Pacing for Multihop Wireless Networks,” *in Proceedings ACM MobiHoc ’05*, pp. 288-299, 2005.

[11]. A. Singh and K. Kankipati, “TCP-ADA: TCP with Adaptive Delayed Acknowledgement for Mobile Ad Hoc Networks,” *in Proceedings IEEE Wireless Communication and Networking Conference (WCNC ’04)*, no. 1, pp. 1679-1684, 2004.

[12]. R.Oliveira and T. Braun, “A Dynamic Adaptive Acknowledgement Strategy for TCP over Multihop Wireless Networks,” *Proc. IEEE INFOCOM ’05*, pp 1863-1874, 2005.

[13]. E. Altman and T. Jimenez, “Novel Delayed ACK Techniques for Improving TCP Performance in Multihop Wireless Networks,” *in Proceedings Eighth International Conference on Personal Wireless Communication (PWC ’03)*, pp. 237-253, 2003.

[14]. K. Chen, Y. Xue, and K. Nahrstedt, “On Setting TCP’s Congestion window Limit in Mobile Ad Hoc Networks,” *Wireless Communication. and Mobile Computing*, vol. 2, no. 1, pp. 85-100, 2002.

Optimized Fuzzy Logic Based Segmentation for Abnormal MRI Brain Images Analysis

Indah Soesanti[1], Adhi Susanto[2], Thomas Sri Widodo[2] and Maesadji Tjokronagoro[3]

[1] Department of Electrical Engineering and Information Technology, Gadjah Mada University, Yogyakarta, Indonesia

[2] Department of Electrical Engineering and Information Technology, Gadjah Mada University, Yogyakarta, Indonesia

[3] Faculty of Medicine, Gadjah Mada University, Yogyakarta, Indonesia

Abstract

In this paper an optimized fuzzy logic based segmentation for abnormal MRI brain images analysis is presented. A conventional fuzzy c-means (FCM) technique does not use the spatial information in the image. In this research, we use a FCM algorithm that incorporates spatial information into the membership function for clustering. The FCM algorithm that incorporates spatial information into the membership function is used for clustering, while a conventional FCM algorithm does not fully utilize the spatial information in the image.The advantage of the technique is less sensitive to noise than the others. Originality of this research is focused in application of the technique on a normal and a glioma MRI brain images, and analysis of the area of abnormal mass from segmented images. The results show that the method effectively segmented MRI brain images, and the segmented normal and glioma MRI brain images can be analyzed for diagnosis purpose. The area of abnormal mass is identified from 7.15 to 19.41 cm^2.

Keywords: *Adaptive image segmentation, FCM clustering, abnormal MRI brain image, fuzzy membership function.*

1. Introduction

Image segmentation is one of the most important step to extract information in image processing. Segmentation has wide application in medicine area. The main objective of segmentation of medical image is to partition the image into mutually exclusive and exhausted regions such that each region of interest is spatially contiguous and the pixels within the region are homogeneous with respect to a predefined criterion. Magnetic Resonance Imaging (MRI) is the state-of-the-art medical imaging technology which allows cross sectional view of the body with unprecedented tissue contrast. MRI provides a digital representation of tissue characteristic that can be obtained in any tissue plane. The images produced by an MRI scanner are best described as slices through the brain. MRI has the added advantage of being able to produce images which slice through the brain in both horizontal and vertical planes. The objective of segmenting different types of soft-tissue in MRI brain images is to label complex structures with complicated shapes, as white matter, grey matter, CSF and other types of tissues in neurological conditions.

A variety of segmentation methods have been developed to satisfy increasing requirement of image segmentation over past several years. FCM (Fuzzy c-means) is unsupervised technique that has been successfully applied to future analysis, clustering, and classifier designs in the fields such as astronomy, geology, medical imaging, target recognition, and image segmentation. An image can be represented in various feature spaces, and the FCM method classifies the image by grouping similar data points in the feature space into clusters.

During the past many researchers in the field of medical imaging and soft computing have made significant survey in the field of image segmentation. There has been considerable interest recently in the use of segmentation methods based on fuzzy logic, which retain more information from the original medical image than hard segmentation methods (e.g. Bezdek et al. [1], Udupa et al. [2], Pham [3], Masoole and Moosavi [4]). The fuzzy C-means (FCM) clustering, in particular, can be used to obtain a accurate segmentation via fuzzy pixel classification. Unlike hard clustering methods which force pixels to belong exclusively to one class, FCM allows pixels to belong to multiple classes with varying degrees of membership functions. This approach allows additional flexibility in many applications and has recently been used in processing of medical images [5]. For example, in their segmentation of MRI brain images, Pham et al. [5] thresholded the FCM memberships in order to extract pixels which a high confidence of correct classification. Xu et al. [6] used deformable surfaces that converged to the peaks of the memberships. The FCM method, however, does not address the intensity inhomogeneity artifact that occurs in nearly all MRI [7]-[8]. Originality of this research is the method applied on a normal MRI brain image and a glioma MRI brain images, and analyze the area of abnormal mass from segmented images.

2. Fundamental Theory

2.1 Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI) is an imaging used primarily in medical settings to produce high quality medical images of the soft tissues. MRI is an imaging technique used primarily in medical settings to produce high quality images of the inside of the human body. In this section we give a brief description of the principles of MRI which are refered to [9]. In MRI, the image is a map of the local transverse magnetization of the hydrogen nuclei. This transverse magnetization in turn depends on several intrinsic properties of the tissue. MRI is based on the principles of nuclear magnetic resonance (NMR). The NMR phenomenon relies on the fundamental property that protons and neutrons that make up a nucleus possess an intrinsic angular momentum called spin. When protons and neutrons combine to form nucleus, they combine with oppositely oriented spins. Thus, nuclei with an even number of protons and neutrons have no net spin, whereas nuclei with an odd number of protons or neutrons possess a net spin. Hydrogen nuclei have an NMR signal since its nucleus is made up of only a single proton and possess a net spin. The human body is primarily fat and water, which have many hydrogen atoms. Medical MRI primarily images the NMR signal from the hydrogen nuclei in the body tissues.

The net spin of the nucleus around its axis gives it an angular moment. Since the proton is a positive charge, a current loop perpendicular to the rotation axis is also created, and as a result the proton generates a magnetic field. The joint effect of the angular moment and the self generated magnetic field gives the proton a magnetic dipole moment parallel to the rotation axis. Under normal condition, one will not experience any net magnetic field from the volume since the magnetic dipole moments are oriented randomly and on average equalize one another.

When placed in a magnetic field, a proton with its magnetic dipole moment processes around the field axis. The frequency of this precession, v_0, is the resonant frequency of NMR and is called the Larmor frequency. The precession frequency is directly proportional to the strength of the magnetic field, i.e.

$$v_0 = \mathrm{g}B_0 \qquad (1)$$

where B_0 is the main magnetic field strength, and g is a constant called gyromagnetic ratio which is different for each nucleus (42.56 MHz/Tesla for protons).

Given a specimen, the application of a magnetic field B_0 would create a net equilibrium magnetization M_0 per cubic centimeter, which is aligned to the B_0 field. The M_0 is the net result of summing up the magnetic fields due to each of the H nuclei and is directly proportional to the local proton density (or spin density). However, M_0 is many orders of magnitude weaker than B_0 and is not directly observable. By tipping M_0 away from the B_0 field axis with an appropriate RF pulse having a frequency equals to the Larmor frequency, a longitudinal magnetization component *ML* and a transverse magnetization component *MT* is produced. When the RF pulse is turned off, the longitudinal magnetization component *ML* recovers to M_0 with a relaxation time *T1*, and the transverse magnetization component *MT* dephases and decays to zero with a relaxation time *T2* 1. During relaxation, the protons lose energy by emitting their own RF signal with the amplitude proportional to *MT*. which is referred to as the *free-induction decay* (FID) response signal. *T2* indicates the time constant required from a given tissue type to decay for the FID response signal. The FID response signal is measured by an RF coil placed around the object being imaged.

In MR imaging, the RF pulse is repeated at a predetermined rate. The *repetition time*, *TR*, is the period of the RF pulse sequence is. The FID response signals can be observed at various times within the *TR* interval. *echo delay time*, *TE* is the time between which the RF pulse is applied and the response signal is measured. The *TE* is the time when the spin echo occurs due to the refocusing effects of the 180 degree refocusing pulse applied after a delay of *TE*/2 from the RF pulse. The *TR* and *TE* control how strongly the local tissue relaxation times, *T1* and *T2*, affect the signal. By adjusting *TR* and *TE* the acquired MR image can be made to contrast different tissue types.

2.2 Image Segmentation

The aim of medical image segmentation on human graphical interaction is to define regions, using methods such as manual slice editing, region painting and interactive thresholding. Rajapakse [13] had classified the different methods of image segmentation as four categories. (1) Thresholding region growing and edge based techniques. (2) The maximum-likelihood-classifier (MLC). These methods are basically depend on the prior model and its parameters. Vannier et al. [14] reported satisfactory preliminary results with Bayesian MLC. Ozkan et al. [15] researched about the MLC and the neural network classifier which showed the superiority of the neural network.

Some new methods of image segmentation that could be classified as statistical methods have been introduced in the past few years. Hansen [16] used a probabilistic supervised relaxation technique for segmenting 3D medical images. The method introduced the use of cues to guide the medical image segmentation. Those cues marked by the user have the mean and standard deviation as description parameters. (3) The neural networks methods one example of which is the work of Ahmed et al. [14] who used a two stages neural network system for CT/MRI image segmentation. The first stage is a self-organized principal component analysis (SOPCA) network and the second stage consists of a self-organizing feature map (SOFM). The results obtained compare favorably with the classical and statistical methods. (4) The Fuzzy Clustering methods. In [18] a comparison between the fuzzy logic and artificial neural network techniques in segmenting magnetic resonance images debated for the need of unsupervised technique in segmentation which was provided using the unsupervised fuzzy c-mean algorithm.

3. Fuzzy C-Means Clustering

The FCM method assigns pixels to each cluster by using fuzzy membership functions. Let X=(x_1, x_2,.,x_N) denotes an image with N pixels to be partitioned into c clusters, where x_i represents multispectral (features) data. The algorithm is an iterative optimization that minimizes the cost function defined as follows [16]:

$$J = \sum_{j=1}^{N}\sum_{i=1}^{c} u_{ij}^{m} \| x_j - v_i \|^2 \qquad (2)$$

where u_{ij} represents the membership function of pixel x_j in the ith cluster, v_i is the ith cluster center, and m is a constant. The parameter m controls the fuzziness of the resulting partition, and m = 2 is used in this study.

The cost function is minimized when pixel close to the centroid of their clusters are assigned high membership values, and low membership values are assigned to pixels with data far from the centroid. The fuzzy membership function represents the probability that a pixel belongs to a specific class. In the FCM algorithm, the probability is dependent solely on the distance between the pixel and each individual cluster center in the feature domain. The membership functions and cluster centers are updated by the following:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{\| x_j - v_i \|}{\| x_j - v_k \|}\right)^{2/(m-1)}} \qquad (3)$$

and

$$v_i = \frac{\sum_{j=1}^{N} u_{ij}^{m} x_j}{\sum_{j=1}^{N} u_{ij}^{m}} \qquad (4)$$

where $u_{ij} \in [0, 1]$.

Starting with an initial guess for each cluster center, the FCM converges to a solution for v_i representing the local minimum or a saddle point of the cost function. Convergence can be detected by comparing the changes in the membership function or the cluster center at two successive iteration steps.

One of the important characteristics of an image is that neighboring pixels possess similar feature values, and the probability that they belong to the same cluster is great. This spatial relationship is important, but it is not utilized in a konvensional FCM algorithm. To exploit the spatial information, a spatial function is defined as.

$$h_{ij} = \sum_{k \in NB(x_j)} u_{ik} \qquad (5)$$

In this formula, NB(x_j) represents a square window centered on pixel x_j in the spatial domain. A 3 x 3 window was used throughout this work. Just like the membership function, the spatial function h_{ij} represents the probability that pixel x_j belongs to *i*th cluster. The spatial function of a pixel for a cluster is large if the majority of its neighborhood belongs to the same cluster. The spatial function in incorporated into membership function as follows:

$$u'_{ij} = \frac{u_{ij}^{p} h_{ij}^{q}}{\sum_{k=1}^{c} u_{kj}^{p} h_{kj}^{q}} \qquad (6)$$

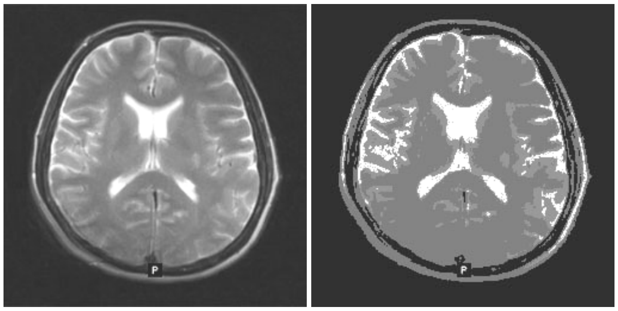
In this formula, p and q are parameters to control the relative importance of both functions. In a homogenous region, the spatial functions fortify the original membership, and the clustering result remains unchanged. However, for a noisy pixel, this formula reduces the weighting of a noisy cluster by the labels of its neighboring pixels. As a result, misclassified pixels from noisy regions or spurious blobs can easily be corrected. The spatial FCM with parameter p and q is denoted $sFCM_{p,q}$. Note that $sFCM_{1,0}$ is identical to the conventional FCM.

The clustering is a two-pass process at each iteration. The first pass is the same as that in standard FCM to calculate the membership function in the spectral domain. In the second pass, the membership information of each pixel is mapped to the spatial domain, and the spatial function is computed from that. The FCM iteration proceeds with the new membership that is incorporated with the spatial function. The iteration is stopped when the maximum difference between two cluster centers at two successive iterations is less than a threshold (ε=0,02). After the convergence, defuzzification is applied to assign each pixel to a specific cluster for which the membership is maximal

4. Experimental Results

4.1 Normal MRI Brain Image

In this research, MRI image is segmented using FCM method incorporated into the spatial information. Figure 1(a) shows the 256x256 grayscale original T2-weighted MRI1 brain image. Fig. 1(b) shows the result of the FCM incorporated into the spatial information with parameters p=1 and q=2.



(a) Original MRI1 Brain image (b) Segmented MRI1 Brain image

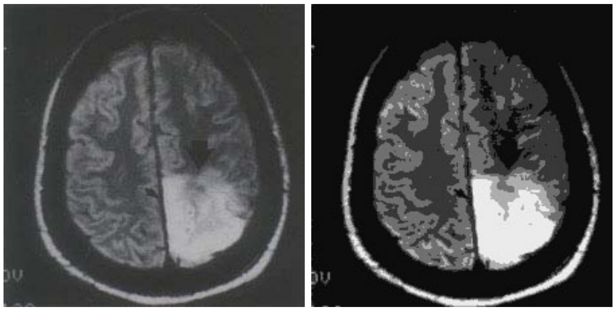
Figure 1. MRI1 Brain image

As can be seen in the segmented images in Fig. 2, lesion or abnormal mass is not identified, and the ventricular system is not extensive and it is a median. So, the image is normal brain MRI image.

4.2 Glioma MRI Brain Images

Figure 2(a) shows the 256x256 original MRI2 image [17]. As can be seen in the segmented image in Fig. 2(b), there is extensive edema extending anteriorly and posteriorly and involving the basal ganglia.

The shift and mass effect on the ventricle have resulted in compromise of the foramen of Monro, and there is evidence of active hydrocephalus, as shown by both the ventricular enlargement and a homogeneous increase through the periventricular region. An additional area of increased signal, separated from the right hemisphere in the left side, suggest that there may even be a metastatic focus or evidence of distant extension of tumor into the left hemisphere. All these features suggest a very high Grade IV glioma or glioblastoma.



(a) Original MRI2 Brain image (b) Segmented MRI2 Brain image

Figure 2. MRI2 Brain image

4.3 Analysis of the Area of Abnormal Mass

In this study, we also apply the FCM method to segment four 256x256 MRI brain images with abnormal mass (i.e. MRI3, MRI4, MRI5, MRI6, MRI7, and MRI8), as shown in Figure 3 [17].

Application of the FCM method in segmentation of the images is in order to analysis of the area of abnormal mass Figure 4 shows the 256x256 segmented MRI3-MRI8 images.

Table 1 shows area of abnormal mass of segmented MRI images. The results are MRI images identified tumors of 2.66% to 7.22% or 7.15 to 19.41 cm^2. In the brain tumor, glioma, the bigger area of tumor the higher grade of glioma.

Table 1. Area of abnormal mass

Images	Areas in %	Areas in cm^2
MRI3	2.66	7.15
MRI4	3.05	8.20
MRI5	3.57	9.60
MRI6	3.61	9.70

MRI7	6.04	16.24
MRI8	7.22	19.41

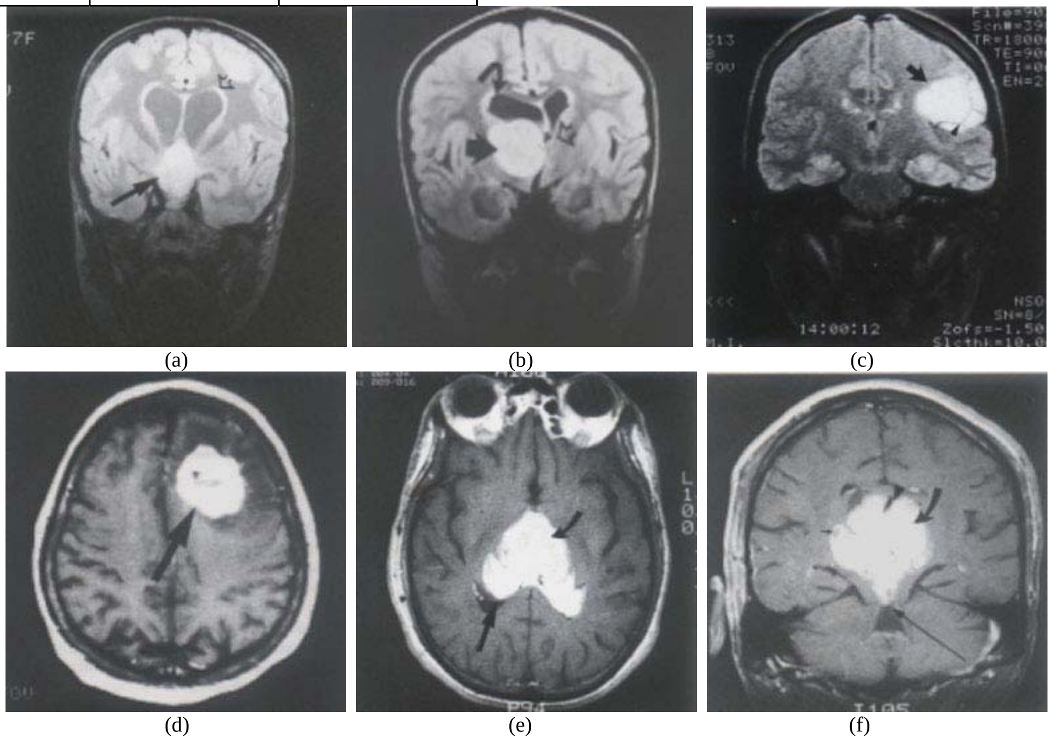


Figure 3. Original abnormal mass MRI brain images: (a) MRI3, (b) MRI4, (c) MRI5, (d) MRI6, (e) MRI7, (f) MRI8.

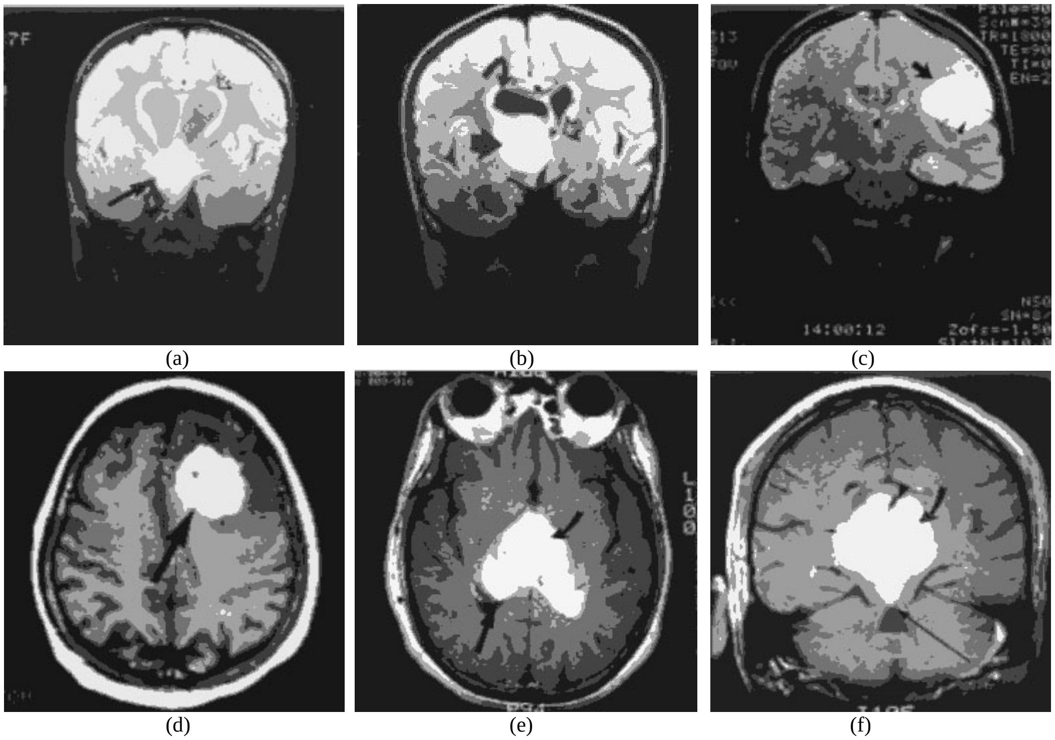


Figure 4.(a) Segmented MRI3, (b) segmented MRI4, (c) segmented MRI5, (d) segmented MRI6, (e) segmented MRI7, (f) segmented MRI8.

5. Conclusions

In this paper we apply an extended FCM method that incorporates the spatial information into the membership function to improve the results of MRI brain image segmentation. The membership functions of the neighbors centered on a pixel of MRI brain image in the spatial domain are enumerated to obtain the cluster distribution statistics. These statistics are transformed into a weighting function and incorporated into the membership function. This neighboring effect reduces the number of spurious blobs and biases the solution toward piecewise homogeneous labeling. The technique was used to analyze a normal MRI brain image and glioma MRI brain images. We applied the technique on a normal MRI brain image and on a glioma MRI brain image. The results show that the method effectively segmented Magnetic Resonance Imaging (MRI) brain images with spatial information, and the segmented normal and glioma MRI brain images can be analyzed for diagnosis purpose. The results are MRI images identified tumors of 7.15 to 19.41 cm^2.

References

[1] J. Bezdek. L. Hall. and L. Clarke. “Review of MR image segmentation using pattern recognition”. Medical Physics. vol. 20. 1993. pp. 1033–48.

[2] J. K. Udupa. L. Wei. S. Samarasekera. Y. Miki. M. A. van Buchem. and R. I. Grossman. "Multiple sclerosis lesion quantification using fuzzy-connectedness principles." IEEE Transactions on Medical Imaging. vol. 16. 1997. pp. 598-609.

[3] D.L. Pham. “Unsupervised Tissue Classification in Medical Images using Edge-Adaptive Clustering”. Proceedings of the 25th Annual International Conference of the IEEE EMBS. Cancun. Mexico. Sep. 17-21. 2003.

[4] L. Jiang and W. Yang. “A Modified Fuzzy C-Means Algorithm for Segmentation of Magnetic Resonance Images”. *Proc. VIIth Digital Image Computing: Techniques and Applications*. Sydney 10-12 Dec. 2003.

[5] D.L. Pham and J.l. Prince. “Adaptive fuzzy segmentation of magnetic resonance images”. IEEE Trans. in Medical Imaging. 1999. Vol. 18. pp. 737–752.

[6] C. Xu. D.L Pham. and J.L. Prince. “Finding the brain cortex using fuzzy segmentation. isosurfaces. and deformable surfaces”. In Proc. XVth Int. Conf. on Inform. Processing in Medical Imaging. (IPMI 1997). pp. 399-404.

[7] S.R. Kannan. “Segmentation of MRI Using New Unsupervised Fuzzy C Mean Algorithm. ICGST-GVIP Journal. Vol. 5. No.2. Jan. 2005.

[8] S. Alizadeh. M. Ghazanfari. and M. Fathian. “Using Data Mining for Learning and Clustering FCM”. International Journal of Computational Intelligence. Vol. 4. No. 2. Spring 2008.

[9] A. Wee, C. Liew, and H. Yan. “Current Methods in the Automatic Tissue Segmentation of 3D Magnetic Resonance Brain Images”. Current Medical Imaging Reviews, Vol. 2, No. 1, 2006, pp. 1-13.

[10] J.C. Rajapakse. J.N. Giedd. and J.L. Rapoport. “Statistical Approach to Segmentation of Single Channel Cerebral MR Images”. IEEE Trans. on Medical Imaging. Vol.16. No.2. April 1997.

[11] M.W. Vannier. C.M. Speidel. and D.L. Rickman. “Magnetic resonance imaging multi spectral tissue classication”. Journal of NIPS. vol.3. Aug. 1991.

[12] M. Ozkan. B.M. Dawant. and R.J. Maciunas. “Neural Network_Based Segmentation of Multi_Modal Medical Images. A Comparative and Prospective Study”. IEEE Transactions on Medical Imaging. Vol. 12. No.3. Sep. 1993. pp. 534-544.

[13] M.W. Hansen and W.E. Higgins. “Relaxation Methods for Supervised Image Segmentation”. IEEE Trans. on Pattern Analysis and Machine Intelligence. Vol. 19. No. 9. Sep. 1997.

[14] M.N. Ahmed and A.A. Farag. “Two stages Neural Network for Medical Volume Segmentation”. Accepted for Publication in the Journal of Pattern Recognition Letters. 1998.

[15] L.O. Hall. A.M. Bensaid. L.P. Clarke. R.P. Velthuizen. M.S. Silbger. and J.C. Bezdek. “A Comparison of Neural Network and Fuzzy Clustering Techniques in Segmenting Magnetic Resonance Images of the Brain”. IEEE Transactions on Neural Networks. Vol. 3. No. 5. Sep. 1992. pp. 672-681.

[16] K.S. Chuang. H.L. Tzeng. S. Chen. J. Wu. and T.J. Chen. “Fuzzy C-Means Clustering with Spatial Information for Image Segmentation”. Computerized Medical Imaging and Graphics. Vol. 30 (2006). Elsevier. pp. 9–15.

[17] J.H. Besese. Cranial MRI. A Teaching File Approach. McGraw-Hill. International Edition. Medical Series. 1991.

Indah Soesanti is with the Department of Electrical Engineering and Information Technology. Gadjah Mada University. Yogyakarta. Indonesia. She received B.S. and M.Eng. from Gadjah Mada University. Yogyakarta. Indonesia in 1998 and 2001. respectively. He is currently a Ph.D. student at the Electrical Engineering. Gadjah Mada University. She has more than 7 years of experience in teaching. Her research interests include image processing. signal processing. fuzzy logiz and its application. optimization. and information system. She has published more than five papers in national journals. She has also presented more than ten research articles in national and international conferences.

Adhi Susanto is a Professor in the Department of Electrical Engineering and Information Technology. Gadjah Mada University. Yogyakarta. Indonesia. He received M.Sc. and Ph.D. from University of California. Davis. US in 1966 and 1986. respectively. He has more than 40 years of experience in teaching and research. His current area of research includes image processing. signal processing. neural networks. wavelets. fuzzy logiz and its application. optimization and control. instrumentation. and information system. He has published more than ten papers in referred international journals. He has also presented more than twenty research articles in national and international conferences. He has written few books related to his research work.

Thomas Sri Widodo is a Professor in the Department of Electrical Engineering and Information Technology. Gadjah Mada University. Yogyakarta. Indonesia. He received DEA. and Ph.D. from Universite des Sciences et Techniques du Languedoc. Montpellier 2. France in 1986 and 1988. respectively. He has more than 30 years of experience in teaching and research. His current area of research includes electronics. telecommunication. image processing. signal processing. neural networks. wavelets. fuzzy logiz and its application. instrumentation and control. biomedical engineering. and hypertermia. He has published more than ten papers in referred international journals. He has also presented more than fourty research articles in national and international conferences. He has written few books related to his research work. He is currently dealing with few projects sponsored by government of Indonesia (the Ministry of Education and the State Ministry of Research and Technology).

Maesadji Tjokronagoro is a Professor in the Faculty of Medicine. Gadjah Mada University. Yogyakarta. Indonesia. He received M.S. and Ph.D. from Gadjah Mada University. Yogyakarta. Indonesia in 1981 and 1986. respectively. He has more than 30 years of experience in teaching and research. His current area of research includes radiology. radiotherapy. medical image processing. biomedical engineering. and hypertermia.

Intelligent Scheduling in Health Care Domain

Srividya Bhat , Nandini S. Sidnal , Ravi S. Malashetty , Sunilkumar. S. Manvi

1 Dept. of PG Studies, VTU, Belgaum, Karnataka, India

2 Dept. of CSE, K.L.E.S.C.E.T, Belgaum, Karnataka, India

3 Dept. of PG Studies, VTU, Belgaum, Karnataka, India

4 Dept. of ECE, REVA ITM, Bangalore, Karnataka, India

Abstract

Healthcare organizations are facing the challenge of delivering high-quality services through effective process management at all levels-locally, regionally, nationally, and internationally. Patient scheduling becomes an integral part of daily work for healthcare professionals. The presented work is to build an agent based information services for mobile users. An agent is characterized by the concepts of situatedness, autonomy and flexibility. Multi-Agent systems (MAS) are appropriate in many medical domains, due to the characteristics of the problems in this area and are the basis of an emerging technology that promises to make it much easier to design and implement. The paper work integrates accessing distributed health care services in multi-agent environment to achieve better Quality of service by using java platform. This develops a framework to schedule the meeting between the patients and the relevant doctors meeting in an efficient way for routine and emergency services.

Keywords: *Mobile users, Software agents, Multiagent systems, medical ontology, FIPA-ACL.*

1. **Introduction**

Most of the Health care professionals use computer systems to access patient's medical record or information about hospital resources and to fix an appointment for multiple patients with potentially conflicting schedules. Meeting Scheduler in health care domain is considered as a part which will grow most rapidly and lead to economical and popular methodology with autonomous agents, which can schedule meetings and manage calendars on behalf of their users by saving the patients and physicians time. Also this system is generally designed to guide remote patients to fix the appointment with doctor through online facilities and help them to reach an appropriate hospital in an unknown city. A brief introduction of the concepts and methods are used to carry out the paper work are given.

At present most of the hospitals follow a Simple GUI based applications to maintain their information regarding the patients, Doctor and scheduling information. Effective and timely communication between patients, physicians, and other healthcare professionals is vital to good healthcare. Current communication mechanisms are based largely on paper records and prescriptions, which are old-fashioned, inefficient, and unreliable. In an age of electronic record keeping and communication, the healthcare industry is still tied to paper documents that are easily misled, often illegible, and easy to forge.

Healthcare professionals working in highly dynamic hospital environments typically have correspondingly dynamic schedules that are difficult to manage. Emergent tasks and shifting priorities result in existing schedules becoming obsolete. Managing patient appointments is an area that typically consumes a great deal of administrative overhead and cost. Clinic and office administrators are typically juggling multiple phone calls, physician requests, and patient demands. It is also a source of frustration for many patients due to the delays and inefficiencies in speaking with the clinic or office administrator. This leads to no-shows, lost revenue, and operational inefficiencies.

An increase in specialization and technology, especially in the health care department requires efficient management of the resources and timely treatment for the patients. Agents are used to solve the patient scheduling problem in the hospitals because they work well in a distributed, decentralized and dynamic environment. An agent is a software program that acts on behalf of a user, typically used to retrieve and process information. An agent is used to represent each patient and resource in the hospitals. Interaction protocols are used to reduce the search space of possible responses to an agent messages.

Multi Agent System (MAS) based Health Care Domain will address some of these issues:

- ❖ MAS contains agents that allow the user to search for medical centers satisfying a given set of requirements, to access his/her medical record or to make a booking to be visited by a particular kind of doctor.

- Some of the agents in the system can provide information about the medical centers that are available in a given city.
- The MAS also contains an agent for each medical center in town; these agents may be asked about the doctors working in that hospital, or may be requested to perform a booking in the schedule of a specific doctor.
- Providing a decomposition of the problem that matched agents to entities which could be realistic players in such a domain and to take care in who had access to which information.

Healthcare professionals working in a hospital environment typically have many responsibilities contending for their time. With tasks ranging from providing medical care and monitoring patients to undertaking administrative responsibilities, it is often the case that healthcare professionals have a seemingly endless set of changing tasks to carry out. Consequently, they must manage their time by composing their activities into prioritized to-do lists. However, hospitals are inherently highly dynamic environments in which task interruptions and delays are commonplace. Additionally, previously unforeseen tasks can emerge that may require attention alongside the already scheduled tasks. In the face of such change, static paper or whiteboard-based to-do lists can become difficult to manage and, in the worst case, obsolete.

Recent advances in embedded sensor and mobile computing technology have given rise to a range of possibilities in pervasive healthcare. Among these is the opportunity to aid healthcare professionals by automatically managing their schedules in the face of significant contextual events that can negatively impact their schedule. When a patient wants to arrange an appointment with a doctor, or a doctor must arrange a visit of a patient with a service, it is required to schedule a meeting according to different constraints such as timetable of services or doctors, and agenda of the patient.

In the proposed scheme, any number of patients can access the scheduling system through patient-agent by filling all the details such as nature of disease, preferable time and date provided in the meeting request form to fix the appointments with appropriate doctor by searching the nearby hospital in the city. Upon receiving the patient request doctor agent will accordingly schedule, reschedule, or postpone the appointment meetings by viewing the available date in the doctor appointment calendar. If the patient arrival occurs at emergency case, the doctor agent will give first preference to emergency case and reschedule the appointment of already scheduled meetings and convey the same to the concerned patient. This scheduling system will reduces the conflicts between patients by negotiating best available date for meeting.

1.1 Why Agents?

Before answering the question of why agents might be useful, a few words should be said about what an agent is. Although there is no universal agreement, a popular definition, from [Wool95], describes an agent as a software entity that has the characteristics of, autonomy (acts independently), proactivity (goal-based), reactivity (responds in a timely fashion to events) and social ability (communicates with other agents to achieve goals collaboratively). Other characteristics frequently quoted include mobility (the ability to move from one host to another) and learning (the ability to improve performance overtime based on previous experiences).
Software with the above characteristics offers the possibility of systems which can lower the cost and improve the performance of businesses operations by

- Automating mundane tasks,
- Enabling users/customers to complete tasks that would otherwise be very difficult, time consuming, costly or just impossible, and
- Adapting to unexpected events or changes in the environment automatically.

Of course it may be possible to achieve cost saving and performance boosting solutions without agents, but agent technology provides a more natural model of the real world (i.e. a community of entities each with their own goals, communicating and often working together to achieve mutual benefit) compared to existing software paradigms, such as object-orientation.
Furthermore agent technology consolidates and builds upon a number of important computing technologies (object-orientation, distributed computing, parallel processing, mobile code, symbolic processing) and research results from other disciplines (artificial intelligence, biology, mathematics). In this way, agent technology offers a way to unify and simplify the use of the wide range of software technologies available today.

1.2 Problem Statement

Multi-agent systems are widely used to address large-scale distributed combinatorial real world problems. One such problem is meeting scheduling (MS) in health care domain that is characterized essentially by two features defined from both its inherently distributed and dynamic nature i.e. the presence of patient’s preferences that turn it into a search for an optimal rather than a feasible solution. In this connection at least the following questions arise:

- When should the meeting take place?
- How to reach an appropriate hospital?
- What are the services available within hospital?
- How fast the doctor is available?
- Which Doctor is free to fix an appointment?

- How many patients will meet the doctor in a day, and who are they?

To solve it, today heuristics are used, because there is no optimal algorithm that fits for all possible solutions. The techniques of artificial intelligence are also used. An intelligent agent means that the agent has the knowledge about the interest and priorities of persons. Routine activities of physicians with regard to the meeting scheduling are practiced by agents in that way, that it filters and administrates information and answers questions. Supposing that every patient has got his own calendar, which is administrated by an agent, the reliability of his/her calendar will be very well. Also, a certain security of the private data is guaranteed.

“The problem is to develop a framework for distributed health care services using multi agent systems and to develop and implement an algorithm for the application of intelligent scheduling in health care domain using JAVA technology”.

1.3 Scope of the study

The paper envisages development of framework and demonstration of the feasibility of distributed health care services using cooperating multi agents. Therefore a complex application of scheduling meeting for ‘n’ number of patients has been used. The development of parallel algorithms or task graphs for computations does not lie within scope of this paper.

1.4 Related Works

Here we present accessing of health-care related services by deploying intelligent agents. The software-agent paradigm [3] [4] was adopted due to its autonomous, reactive and/or proactive nature, which comprises of important features in real-time application deployment for dynamic systems like the one under consideration. Furthermore, software agents can incorporate coordination strategies, thus enabling them to operate in distributed environments and perform complex tasks. Software-agent technology is considered an ideal platform for providing data sharing, personalized services, and pooled knowledge. The work in [7] presents the Foundation for Intelligent Physical Agents (FIPA) that defines standards for agent interoperation. The aim in the Agent Cities is the construction of a worldwide publicly accessible network of FIPA based agent platforms. Each platform will support agents that offer services similar to those that can be found in a real city. Once the initial services have been deployed, it will be possible to implement intelligent complex compound services.

In the research literature, there are several agent-based applications reported in the healthcare domain. In particular, one of the earliest examples of work examining the role of multi-agent systems in healthcare is offered by [6]. The focus of the work presented there, and of the broader context, in which it was conducted, is upon appropriate theorem proving in decision support systems that have to deal with complex, incomplete, inconsistent and potentially conflicting data. The agent component is designed to support of tasks amongst players in the system. Heine et al [8] simulate an agent oriented environment for German hospitals with the objective to improve or optimize the appointment scheduling system, resource allocation and cost benefit of clinical trials. Nealon and Moreno [10] have discussed the potential and application of agents to assist in a wide range of activities in health care environments. Mabry et al [9] employ the Multi agent system for providing diagnosis and advice to health care personnel dealing with traumatized patients. Nealon and Moreno [2] have discussed various applications of MAS in health care e.g., coordination of organ transplants among Spanish hospitals, patient scheduling, senior citizen care etc. A research paper, called PalliaSys is offered by [15]. It incorporates information technology and multi-agent systems to improve the care given to palliative patients. An Intelligent Healthcare Knowledge Assistant [12] was developed which uses multi agent system for dynamic knowledge gathering, filtering, adaptation and acquisition from Health care Enterprise Memory unit.

However, it is observed from literature survey that when the Agent Cities initiative was made public, the potential development of agents that could offer not the usual leisure-oriented services but health-care related services. The work here describes automation of a multi-agent system that caters to special types of patients or providing assistance to patients for appointments. So, the concept of intelligent agent and mobile technology is used to achieve automation, efficiency, reliability and scalability in devising Health care domain for distributed, decentralized and dynamic environment to treat the patients efficiently by cutting down the time and cost.

2. Agent Technologies

Agents are considered one of the most important paradigms that on the one hand may improve on current methods for conceptualizing, designing, and implementing software systems and on the other hand may be the solution to the legacy software integration problem.

2.1 What is an Agent?

The term ‘agent’ or software agent has found its way into a number of technologies and has been widely used, for example, in artificial intelligence, database, operating system and computer networks literature. Even within the Agent Research Community, there are at least the following variants on the term agent: Mobile Agents, Learning Agents, Autonomous Agents, Planning Agents, Simulation Agents, and Distributed Agents. Although there is no single definition of an agent, all definitions agree that an agent is essentially a special software component that has autonomy that provides an interoperable interface to an arbitrary system and/or behaves like a human agent, working for some clients in pursuit of its own agenda. Even if an agent system can be based on a solitary agent working within an environment and if necessary interacting with its users, usually they consist of multiple agents. Theses multi agent systems (MAS) can model complex systems and introduce the possibility of agents having common or conflicting goals. These agents may interact with each other both indirectly (by acting on the environment) or directly (via communication and negotiation). Agents may decide to cooperate for mutual benefit or may compete to serve their own interests.

In the data processing technology, an agent is software that supports a person, by executing autonomous several processes. Persons can delegate work to agents, instead of doing them on their own. Agents represent human users. The main difference to traditional software is their relative autonomy, which can be explained as a goal-directed, proactive and self-starting behavior. Software agents run continuous and autonomous in a defined environment, together with other agents and processes. Also, agents need:

- **Social ability:** agents communicate with their users, but also with other agents, using special agent-languages.
- **Reactivity:** agents perceive their environment, which can be their owner, other agents, the internet…and they react on different influences.
- **Pro-activity:** agents not only react on signals, but they also do independent actions, to reach a goal.

2.2 Software Agents

The software agents deal with how to do something, hiding details and work from the user who describes to the agent what to do. Agents act as much as possible without human intervention by learning from users' desires and making decisions for the user. Software agents are also dynamic and responsive to a variable environment. Agents ease and quicken the use of complicated systems. The Definition of a software agent is ambiguous but several key concepts are important. Software agents are usually goal-directed processes, which perform tasks autonomously delegated to them. It is situated in, aware of and reacts to its environment. An agent is also capable of cooperating with other agents, human or software, to accomplish tasks or to get new ones. Software agents are desired to be intelligent and mobile. These capabilities offer a new way to build very large heterogeneous applications.

The agent model working at a high level can be described as skills talented in different areas. Task level skills describe what capabilities the agent has for resolving tasks that user has given and how the agent can observe the environment and ways to handle information, for example database queries. Knowledge has the rules that agent follows as it goes on with a task. This is based on the awareness of the environment. This awareness is received by an agent in different ways: the developer has specified it by programming it in the application platform; the user can specify it by answering questions that the agent needs in its task or the agent can learn it from the environment or from the other agents. Communication skills are the agent's capabilities to communicate with other agents and with the user. The most natural way to communicate with humans would be by speech and facial expressions.

2.3 Taxonomy of agents

Agents can be classified by their capabilities and method of implementation as given bellow:

- **Collaborative agents** are autonomous and they communicate with each other. They can learn, but this is not essential. Collaborative agents have their power in the group.
- **Interface agents** intend to work for the user by helping autonomously, observing users habits and imitating them. The user can also instruct interface agents or they can ask for advice from other agents.
- **Mobile agents** are agents on the move. The ideal situation would be that mobile agents are sent to the Internet to do tasks for the user and when they are ready, to return home. Mobile agents also are capable of communicating and in an ideal situation they don't all collect the same information, they ask for it from other agents.
- **Reactive agents** are impulse driven. They react by producing a response to an impulse. These agents can trigger events and are suitable for handling sensor data. A reactive agent does not actually exchange data but more like knowledge.
- **Hybrid agents** are combinations of the above agents. They can be like GOSSIP, is a combination of an information agent and a mobile agent, which goes to the Internet and collects data for the user.

2.4 Intelligent agents

Agents in common are assumed to be intelligent, serving the user autonomously. The intelligence of agents comes from AI research, which has introduced different kinds of techniques like neural networks and genetic algorithms for problem solving and learning.

2.5 Multi agent cooperation

Coordination of the agents in a system is important to get the agents to reach the overall goal. Because of the distributed expertise, there is a need to coordinate everyone to prevent chaos and to make the system more efficient. Usually the different agents work toward a common goal, and therefore there is no conflict between them. The individual agent's objective does not matter, only the overall system. This is what Wooldridge mean about "benevolence assumption". In contrast some agents are self-interested. These types of agents have goals that will be in conflict with other agents. However, they still need to cooperate and it is important to find the best way to cooperate.

Coherence and coordination are two issues that need to be considered to decide how successful a multi agent system is. Coherence is the ability of a system to behave as a unit. Coherence is measured in terms of solution quality, efficiency of resource usage, conceptual clarity of operation, or how well system performance degrades in the presence of uncertainty or failure. Coordination is a process in which agents engage in order to ensure their community acts in a coherent manner. In a perfectly coordinated system agent do not need to bother about others sub-goal while achieving a common goal.

There are several ways different agents can work together to solve problems. Contracting is one solution to coordinate agents to work together. By using the contract net protocol standardized by FIPA (Foundation for Intelligent Physical Agents), the agents can cooperate by sharing tasks. A manager announces the problem to the other agents. As the agents listen to announcements and evaluate them with respect to their own resources, they place a bid if they find a suitable task. Several agents can bid for the same task, and then the manager has to decide from the information of the bid which agent should win the bidding round and then will be awarding the contract.

Another way to let agents cooperate to solve a problem, is result sharing. This will typically be that each agent solve small problems which later on will be become larger solution. Result sharing is when agents may share information relevant to their sub-problems. Durfee has suggested 4 ways to improve group performance:

- **Confidence:** When independently derived solutions can be crosschecked, the confidence in the overall solution is increased.
- **Completeness:** Agents that share their local views to achieve a better overall global view.
- **Precision:** The precision of the overall solution is increased when agents share results.
- **Timeliness:** As several agents work on the solution, the result could be derived more quickly.

2.6 Multi agent Systems

A multi-agent system (MAS) is a system composed of multiple interacting intelligent agents. Multi-agent systems can be used to solve problems which are difficult or impossible for an individual agent or monolithic system to solve. Examples of problems which are appropriate to multi-agent systems research include online trading, disaster response, and modeling social structures. The agents in a multi-agent system have several important characteristics:

- **Autonomy:** the agents are at least partially autonomous i.e. agents operate without direct human intervention and have control over their own actions.
- **Local views:** no agent has a full global view of the system, or the system is too complex for an agent to make practical use of such knowledge
- **Decentralization:** there is no one controlling agent.

3. Proposed Scheme

This section describes the proposed model in terms of the network environment, hospital environment, patient environment and agencies involved in building and maintaining the medical data center, the agent interactions in discovering and building an automated meeting scheduler in health care domain to access distributed health care services, also the advantages and limitations of this proposed system.

3.1 Network Environment

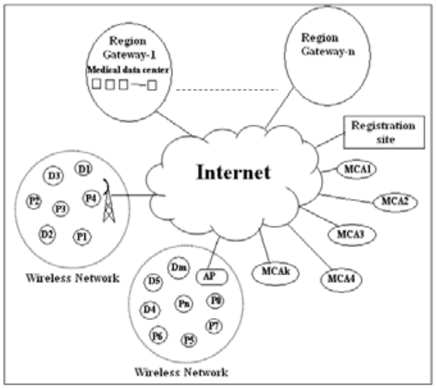


Fig. 1 Network Environment

Network environment for the proposed work is depicted in Fig. 1. The network environment consists of clusters of medical center agents (MCA1…MCAk) in a fixed network, regional gateways, a registration site, mobile patients (P1…Pn) and doctors (D1…Dm) in the wireless environment. Clusters are categorized based on their physical geographical locations where each cluster consists of medical center agents hosting several medical centers. The gateways are connected to the network based on the regions.

Mobile users or patients are in the vicinity of a wireless local area or in a cellular network. The mobile users or patients in a particular region request its regional gateway to fix / schedule an appointment with the doctor. The gateway comprises of medical center data, case base and the patient preferences to identify the relevant medical centers and doctors to coordinate the meeting scheduling process. An agent platform exists in all the components of network environment to facilitate agent based activities, since the information that must be dealt with is geographically distributed. The servers hosting medical centers are reliable and have sufficient bandwidth with good connectivity to accept requests from large number of mobile patients.

3.2 Hospital Environment

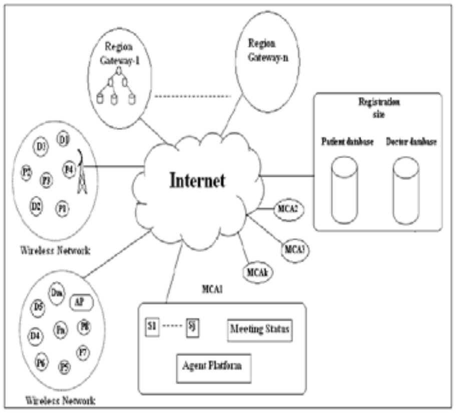


Fig. 2 Hospital Environment

The Fig. 2 depicts the Hospital Environment. The hospital environment consists of a number of Medical center agents that run concurrently on various servers that are connected to the WWW. Each medical center agent will keep its own data, and each doctor will have his/her personal information i.e. an up-to-date daily schedule in a personal computer comprising an agent platform that hosts an agency to carry out the meetings and communication with the patients. Meeting status contains each patients scheduling information such as confirmation of the appointments, postpone or rescheduling of the meetings. The registration site maintains the databases of patients and doctors also the medical records of the potential patients of the system. The medical data center on the regional gateway provides the information about all the medical centers and doctors working in that medical center. The agent platform supports persistence, security, communication and computing services.

3.3 Patient Environment

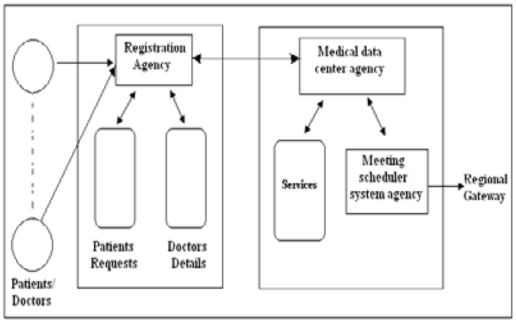


Fig. 3 Patient Environment

The Fig. 3 depicts the Patient Environment. The mobile patients register themselves in the registration desk of the regional gateway to fix an appointment with appropriate doctor by searching the nearby medical centers. The agencies involved in patient environment are registration agency, medical data center agency and meeting scheduler system agency. These agencies employ static and mobile agents to perform the dedicated tasks and focuses on scheduling the meeting for patients and building repository of services for mobile patients based on the patient agenda. Also these agencies will automate the process by enabling the mobile patients to complete the meeting scheduling process successfully without continuous online presence.

3.4 Advantages

The advantages of the proposed systems are as follows:

- ❖ Autonomy of fixing the appointments with doctors as per the patient's requests is achieved.
- ❖ Secure user access to medical records at any time.
- ❖ Support for user queries about the medical centers, and availability of doctors in the medical centers.
- ❖ Online booking for appointments with specialist doctors, whose offices in turn, automatically receive the appropriate medical records for reference and updating.
- ❖ High level accuracy and system reliability.
- ❖ Better time efficiency and flexibility due to quick and efficient retrieval of information any time.

4. Requirement Analysis and Specification

It is widely believed that the next generation of computer desktop applications will be significantly more proactive in helping users to achieve their goals than those which currently exist. Rather than the user having to specify each and every step of a given task, the desktop of the future will be composed of a series of intelligent agents to which a number of high level tasks can be delegated. These agents will be responsible for autonomously deciding how the task is to be achieved and actually performing the necessary set of actions, including handling possible interactions with other intelligent agents. This chapter reports on the requirement analysis and specification of a particular agent-based application which arranges meetings for patients and doctors.

4.1 Problem Statement

To implement an agent based meeting scheduling system, which can schedule meetings for a set of patients. The patient, who wants to schedule a meeting between doctors, just fills the meeting request form that will be provided by an interface. When the patient submits it, the negotiation and scheduling process have to be automatic initiated and its control should be taken by the agent residing in the machine. The agents have to cooperate and do the negotiation on the behalf of each patient. An agent should work for every patient, so that the negotiation is done by it and the participant patients need not be present on their machine, provided their computer machine should be on and running this scheduling system, so the patients can view his updated meeting status any time.

4.2 Functional Requirements

Input: The patient fills all the details provided in the meeting request form to fix the appointments with appropriate doctor and submits it.

Output: The meeting is scheduled on a date and time convenient for all patients taking into account all the patient preferences. And the details of the scheduled meeting are added and displayed in the meeting status of every concerned patient.

5. Design and Implementation

A detailed design using Unified Modeling Language (UML) notation with diagrams and implementation details are given.

5.1 System Design

Designing a system mainly focuses on the detailed implementation of the proposed systems. It emphasizes on translating performance specifications recorded at the time of system study into design specifications. System Design phase is a transition from a user-oriented document to on the methods adopted for developing the system. Design part is the pivotal point in the system development life cycle.

5.1.1 Design

In this phase the architecture of the proposed system is conceived and developed. The architectural diagram helps in a smooth transition between the design stage and the implementation stage. The various factors that are considered before developing the system architecture are cost, reliability, accuracy, security, control, integration, expandability, availability, and acceptability.

The elegant design achieves its objectives with minimum use or resources. The system analyst must have clear understanding of the objectives that the design is aiming to fulfill. There is usually more than one way of achieving a desired set of results.

5.2 Architecture

The architectural diagram for the MAS is shown in Fig. 4. The aim of the MAS is to provide access to the basic health care services in a given city to the patient and to schedule the meeting between patient and doctor. The architecture shows interactions among agents, and also the interactions between humans/resources and agents.

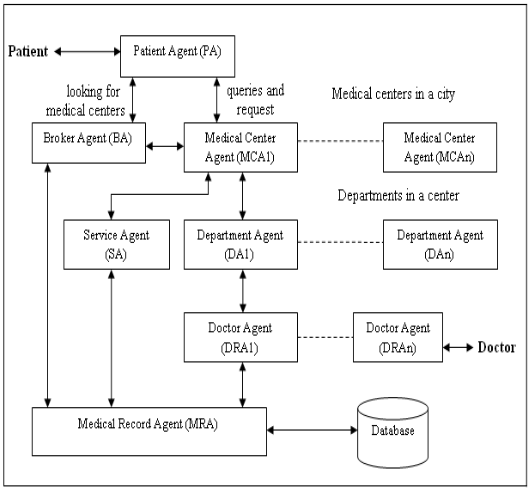


Fig. 4 Architectural diagram for the MAS

A patient interacts with the system through a Patient-agent (PA), provided a GUI through which patients could make queries and receives answers. This agent stores static data related to the patient such as the national healthcare number, name, address, phone number, and information for allowing a secure access to the system (login, password, keys). It also stores dynamic data such as the agenda of the patient. The static data will be used to identify the patient in the system (authentication and ciphering). The agents of the system will exchange required data automatically in each step, *e.g.* a doctor needs to know personal details of a patient before the medical visit, in order to retrieve his/her medical record from a database. The dynamic data is very useful to guide negotiations between any PA and other agents, because a PA can avoid coincidences in those negotiations, *e.g.* if the patient works from 9:00AM to 14:00PM, his agent would arrange meetings during the afternoon and night.

All PAs can talk with a Broker Agent (BA) provided an interface between all the agents internal to the system and the patient-agents. The BA is the bridge between patients and the medical centers, and it is used to discover information about the system. All PAs can ask this agent in order to find medical centers satisfying certain criteria. The BA covers all the medical centers located in a city or nearby area.

Patient can access the system through the Medical Centre Agent (MCA) that centralizes and monitors the outsider's accesses. Each medical center is represented by Medical center Agent which contains all the information related to the medical center such as address, phone number, opening times, location, and so on. A MCA monitors all of its departments, represented by Department Agents (DAs), and a set of general services represented by Service Agents (SAs), such as a blood test service, ambulance etc. Each department is formed by several doctors represented by Doctor Agents (DRA) with specialization, free time and day of doctors.

Database is used to store all patients' medical records which can be accessed through the Medical Record Agent (MRA). This agent provides a secure access to the data using authentication. When a patient wants to arrange an appointment with a doctor, or a doctor must arrange a visit of a patient with a service, it is required to schedule a meeting according to different constraints such as timetable of services or doctors, and agenda of the patient. Here the patient-agent will search nearby hospitals by selecting city or area and category of hospitals also the available services in the hospitals. The patient-agent will then request for the appointment dates with the doctor through the doctor agent. The doctor agent will view the list of appointment request and accordingly it confirms the request or reschedules the appointment date and time as per the free time and day of doctor and confirm the same to the respective patient. If the patient arrival occurs at emergency case, then doctor agent will give first preference to emergency case and reschedule the appointment of already scheduled meetings and convey the same to the concerned patient. This scheduling system will helps in reducing the conflicts between patients by negotiating best available date for meeting.

The goal is to create an automated meeting scheduling agent in health care domain that is:

- ❖ It allows the patients to input his/her meeting request.
- ❖ Negotiates with the agents of the other requested patients.
- ❖ Finds out its best fitting and free time slots.
- ❖ Compares them with the sent fitting slots of the patient-agents and find out the best ones.
- ❖ Reacts to the incoming patient request by sending back its best fitting free time slots.
- ❖ Shows all fixed meeting in a time table.
- ❖ Allows the patient to input his/her preferences.

5.3 Phases

This paper implements the above mentioned goals by using Java programming tools. The phases include the complete life cycle of the multiagent system to schedule the meeting between the patients and doctors are given bellow in steps.

Step 1: Register the members to the centralized database server. This centralized database server maintains a list of hospitals in a city, with each hospital containing different departments with associated services and list of doctors with different specialization and free time and day of doctors.

Step 2: When the application on the centralized server is executed, any number of patients can access the system through patient-agent and send the meeting request form by filling all the details such as nature of disease, preferable time and date to fix the appointments with appropriate doctor by searching the nearby hospital in the city.

Step 3: The patient will first open the login page. If he/she is a new patient then he/she will click new patient and Register. The window will be the registration page of the patient. Once the patient registers he will be activated by the broker agent and can easily login.

Step 4: The patient may request information about all the medical centers available in a particular city. If the patient is aware of a specific medical centre in the area, he/she may request information about the medical services and doctors in that centre. Also it is possible to book a visit to a doctor. In this kind of request the patient-agent has to select the Broker Agent as the recipient of the message. As BA is aware of all Medical Center Agents in town, it will find out which of them satisfy the patient's constraints.

Step 5: Broker Agent will have a predefined user name and password through which he will do various operations and he will insert, update any data from the database based on complaints received from patient-agent. Broker agent will deactivate any member at any time. Also Broker agent will add new area, category of hospital, specialty of hospital and new hospital to the database.

Step 6: The Patient-agents sends a request (REQ) to the MCA through BA. This REQ is forwarded to the department selected by the patient-agent. The Department Agents (DA) will send the REQ to the Doctor Agents. The Doctor Agents (DRA) replies to the request, in which it displays the earliest time in which the doctor has a free slot for making a visit. The patient will view his Meeting status any time such as confirmed, postponed, or rescheduling of the meetings.

Step 7: The Doctor will login any time and view the List of recent appointment request, List of forthcoming appointments and the calendar showing the available dates and time for meeting and scheduled meeting time table.

Step 8: The Doctor Agent will assign the time slots of each day, week with their respective priorities for doctors and fix the meeting using date, priority and time into consideration.

Step 9: If the patient arrival occurs at emergency case, the doctor agent will give first preference to emergency case and find a best fitting time slot for the meeting. Finally, the doctor agent confirms that the schedule of the doctor has been modified, and this confirmation is sent to the patient-agent through department agent and medical center agent.

Step 10: The medical records of the patients are stored in a database called Medical Record Agent (MRA), the access to which is controlled by Database Wrapper (DW).

6. Results

In this section, the simulated results obtained with the proposed work are discussed.

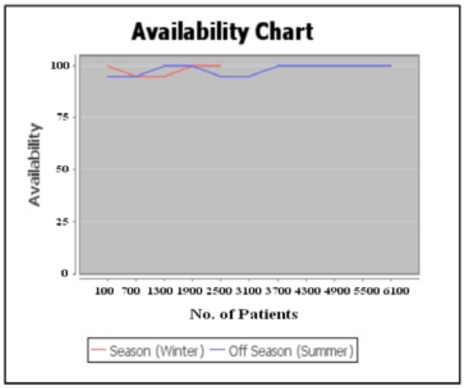


Fig. 5 Success Rate of Getting an Appointment

The Fig. 5 depicts the Success Rate of Getting an Appointment. The X-axis represents the number of patients and Y-axis represents the availability of doctor in terms of percentage during season and off season. During off season, the number of disease will be less (summer) and the availability of doctor will be more when compared to season (winter). During season, the number of diseases will be more, hence the requests will be more and also the availability of doctor will be less.

Season here means - when there are more patients i.e. when there is transition in climate.

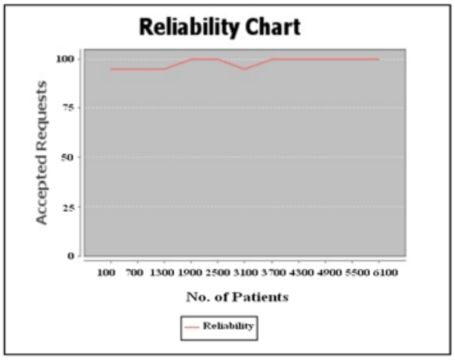


Fig. 6 Reliability of the System

The Fig. 6 depicts the Reliability of the System. The x-axis represents the number of patients and y-axis represents the number of accepted requests in terms of percentage. As the number of patient's increases i.e. the number of requests will be high, so higher the system reliability.

The Figure depicts the Response Time of the System. The X axis represents the number of patients and Y axis represents the time in terms of mili seconds. Response time is the time required to process the request. i.e. after sending the request, how fast it will get confirmed was calculated.

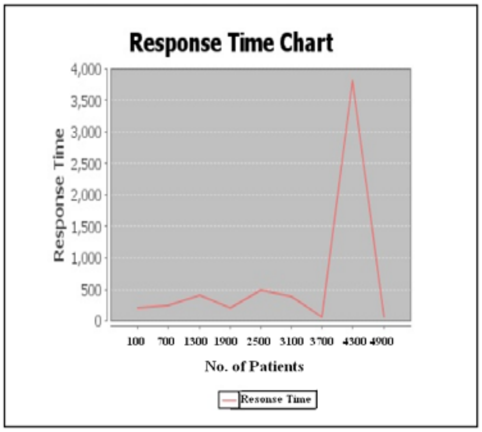


Fig. 7 Response Time of the System

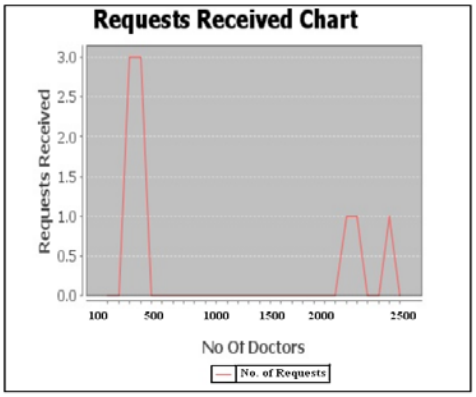


Fig. 8 Success Rate of Getting the Patients

The Fig. 8 depicts the Availability of Patients. The X-axis represents the doctors in the hospitals and Y-axis represents the Number of request received from the patients. More the number of patients, as the number of request received were more.

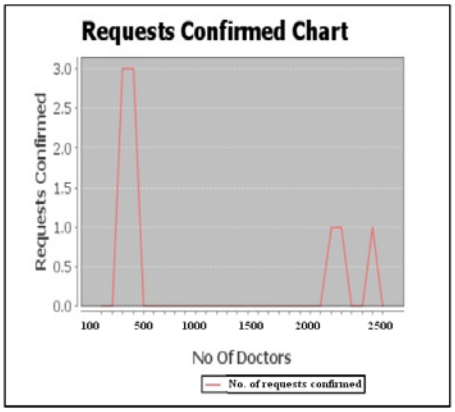


Fig. 9 Reliability of the System

The Fig. 9 depicts the Reliability of the System. The X axis represents the Doctors in the hospitals and Y axis represents the number of accepted requests (processed or confirmed request). As the number of requests increases, the number of requests accepted by the doctor will decrease.

7. Conclusion and Future Work

This paper presented a framework of Intelligent Scheduling in Health Care Domain. The use of agents in health care has experimented an important growth. One of the main benefits of this paradigm is to allow the interoperability of preexisting systems for improving its general performance. We have designed and implemented an agent-based information services for mobile users. The architecture defines the interaction between agents, also between humans and agents. The interaction human-agent is made through personal agents that could be located in computers or mobile devices.

The characteristics of the agent such as the concepts of situatedness, autonomy and flexibility will helps in solving many problems that appear in the health care domain. One such problem we discussed here is access to distributed medical information of a city to schedule the meeting for patients and relevant doctors meeting in an efficient way for routine and emergency services.

Multi agent system was developed to represent the real conditions, courses, and the human decision behavior and to present the overall design of the proposed MAS, emphasizing its architecture and the behavior of each agent of the model, as well as on the scheduling model which provide the activity scheduling process of care and the agent interaction protocol to ensure cooperation between agents that perform coordination tasks for the users. The system implements services as reusable as possible also, it could easily allow the addition of new agents or features to further improve the time efficiency.

In the current implementation the users/patients personal assistant is simulated through a web interface as discussed. In this prototype all the agents are running in the same computer; in order to be usable in a real mobile environment, WAP-accessible version of the MAS must be used. The latest version of JADE, JADE-LEAP v2.4 can be used to achieve better performance and reliability than the existing mechanism.

8. References

1. Zgaya, H. Design and distributed optimization of an information system to aid urban mobility: A multiagent approach to research and composition of services related to transportation. Doctoral Thesis, Ecole Centrale of Lille, 2007.

2. Nealon, J. and Moreno, A. The application of agent technology to health care, Proceedings of the Workshop AgentCities: Research in Large-scale Open Agent Environments, in the 1st International Joint Conference on Autonomous Agents and Multi- Agent Systems (AAMAS „02), pages 169-73, Bologna, Italy, 2002.

3. Weiss, G.: Multiagent systems. A modern approach to Distributed Artificial Intelligence. M.I.T. Press (1999).

4. M. Becker, C. Heine, R. Herrler, and K.-H. Krempels. OntHoS – an ontology for hospital scenarios. In John L. Nealon and Antonio Moreno, editors, Applications of Software Agent Technology in the Health Care Domain, Whitestein Series in Software Agent Technologies, pages 87–104. Birkh¨auser Verlag, Basel, Switzerland, 2003.

5. R. Haux, E. Ammenwerth, W. Herzog, and P. Knaup. Health care in the information society. A prognosis for the year 2013. International Journal of Medical Informatics, 66:3–21, 2002.

6. V. Shankararaman, V. Ambrosiadou, T. Panchal, and B. Robinson. Agents in health care. In V. Shankararaman, editor, Workshop on Autonomous Agents in Health Care, pages

7. FIPA Agent Communication Language: FIPA ACL Message Structure Specification

8. Heine, C., Herrler, R., and Stefan, K. Agentbased Optimisation and Management of Clinical Processes. Proceedings of the 16th European Conference on Artificial Intelligence (ECAI"4)-The 2nd Workshop on Agents Applied in Health Care. 2004.

9. Mabry, Susan L., Hug, Caleb R., Roundy, Russell C. Clinical Decision Support with IM-Agents and ERMA Multi-agents .cbms.page.242. 17th IEEE Symposium on Computer-Based Medical Systems (CBMS"04), 2004.

10. Nealon, J. and Moreno, A. Agent-Based Applications in Health Care. In Applications of Software agent technology in the health care domain. Whitestein Series

in Software Agent Technologies. Birkhauser Verlag, Basel. 2003.

11. Riano, D., Prado, S., Pascual, A. and Martin, S. June. A Multi-Agent System to Support Palliative Care Units. Proceedings of the 15th IEEE Symposium on Computer-Based Medical Systems (CBMS 2002). 2002.

12. Hashmi, ZI., Abidi, SSR. and Cheah, YN.. An Intelligent Agent-Based Knowledge Broker for Enterprise wide Healthcare Knowledge Procurement, Proceedings of the 15th IEEE Symposium on Computer-Based Medical Systems (CBMS 2002). June 2002.

13. Petrie, C. Agent-based software engineering. In: Agent-Oriented Software Engineering. Lecture Notes in Artificial Intelligence 1957, (2001), 58-76.

14. Jennings, N. On agent-based software engineering. Artificial Intelligence 117, (2000), 277- 296.

15. Riano, D., Prado, S., Pascual, A. and Martin, S. June. A Multi-Agent System to Support Palliative Care Units. Proceedings of the 15th IEEE Symposium on Computer Based Medical Systems (CBMS 2002). 2002

Robust Model for Text Extraction from Complex Video Inputs Based on SUSAN Contour Detection and Fuzzy C Means Clustering

Narasimha Murthy K N[1], Dr. Y S Kumaraswamy[2],
[1] Professor, Dept of Information Science and Engineering, V T U, City Engineering College,
Bangalore, Karnataka, Pin-560062, India,

[2] Professor & HOD, Dept. of MAC (VTU), V T U, Dayananda Sagar College of Engineering,
Bangalore, Karnataka, Pin-560078, India,

Abstract:

The proposed system introduces a novel approach for extracting text effectively from different types of complex video inputs. The valuable information within the text can be deployed for text indexing and localization. The proposed system uses contour based protocol like SUSAN algorithm for evaluating the contour detection. The system then explores candidate text area and refines the edges by Fuzzy C Means Clustering. The unwanted non-text portions are removed using morphological operation like dilation. The results obtained from the proposed implementation are then compared with the traditional algorithm used in prior research work for evaluating its efficiency. The result achieved outperforms all the prior algorithms for extracting text from different types of complex video input.

Keyword: *Text Extraction, SUSAN, Fuzzy C Means Clustering, Morphological Operations*

1. Introduction

With the commercial increase in the videos on the various networking and multimedia applications, various service providers are showing increasing attention towards archiving the digital contents for various value added services[1]. The presence of text in video frame is very precious source of high-level semantics and content understanding. For achieving this purpose, text localization and extraction from the videos is the mostly considered option. According to the prior work, text extraction method can be classified into connected component-based approach [2], boundary based approach [3-7], and texture based approach [8-12]. However, component based approach is not so result oriented as it doesn't operated optimally for any video images because of its assumptions that text image elements in the same area will have equivalent color as well as gray-scale strength [2]. The boundary based approach will require the text to be logically elevated contrast for identifying boundaries [3-7]. The texture based approach frequently uses FFT, DCT, wavelets, and Gabor filter for extracting features [13]. Unfortunately, such methods requires high amount of training which is definitely not cost-effective for large scale deputation. Majority of such techniques has also yielded to high rate of false positives results, which motivate the researcher to work on much more efficient algorithm with most challenging deployment environment with highest accuracy and less false positives in the area of text-extraction from videos. Moreover, extracting text from video with higher accuracy is much more challenging in comparison to text extraction from the images, because of diversified problems yielding from complex background, fast moving text in videos, videos with multiple language in one instant, contrast, etc.

Owing to the above mentioned issues, the proposed system will attempt to design a framework for extracting text using SUSAN algorithm for contour based detection, morphological operations like dilation, and Fuzzy C-Means Clustering for much accurate results. Finally the results obtained from the proposed system has been compared with the 5 significant prior research work for establishing the fact that the proposed system has outperformed the existing algorithms for text extractions. We discuss related work in Section II. The research methodology is discussed in Section-III. Proposed system is elaborated in Section IV. Implementation and Results is described in Section-V. Performance Analysis of the proposed system is discussed in Section-VI and finally conclusion and future work is described in Section-VII

2. Related Work

Pratheeba e.t. al [14] has proposed a unique technique for text localization and extraction from complex video input based on findings that there persist colors of contrast nature between text and its adjacent background.

Ghosh e.t. al [15] has proposed an analytical architecture for automated monitoring of news videos with multiple languages. The system combines the audio and visual charecteristics for identifying keywords which characterize a specific news.

Abburu [16] has proposed and analyzed DLER tool for integrating the text identification, localization, extraction, and the recognition method in a single tool

Yen e.t. al [17] has highlighted an effective text extraction algorithm using news video using the temporal information of the video and logical AND operation for removing the most irrelevant background

Qiujun [18] has described a technique for extracting news contents from the web pages based on various non-complex charecteristics seen in majority of the frequently visited websites. The significant feature is the similarity of the dual pages which are gathered from the equivalent topic of a website and published on the same date.

Vijayakumar [19] has proposed an effective text extraction algorithm from sports video. The system can only identify the text in video in the edge of the image. The author has used key frames from the video by color histogram procedure for minimizing the quantity of the video frames.

Stefanos e.t. al. [20] have introduces novel implicit interest indicator for video search and described a new procedure to designing a content similarity graph based on the implicit indicators of patterns of user interaction using SVM classifier.

3. Research Methodology

The proposed system describes a robust text localization and extraction methods from input real-time video with complex background. The text localization and extraction of the proposed system will consists of video frame identification and extraction followed by gray-scale conversion, video segmentation, binarization, contour detection and finally text region extraction. The input from running video will consists of several video frames which will be extracted. Here the video text will be classified into two types: firstly, the type of text which will not change much with the running of the video (e.g. channel name, news headings, date, time etc) and secondly, the type of text which will randomly change in very second as a matter of news update. The situation very complex, when the sports video or news video is considered. The various types of the scene text will be visible only within the frame showing all the scenes like text on hoarding, streets, shops, accessories, vehicles etc. Such types of the text will be subsidiary to the contents on the scene and will be worthy only in certain cases of applications using monitoring or scanning text visible on our known objects. This cannot be use in general text indexing and recovery. Such types of presence of text will be very challenging towards effective text extraction process as it might be observed in infinite quantity with various orientation and structures. Therefore, if such types of occurrences of complex text can be involuntarily identified and extracted, it will lead to evolution to a very higher level of text extraction algorithm which will have higher scope depending on the usage. The need of researching in this types of text extraction is of very high priority as the text extraction tool will effectively be able to capture the current contents of the text without any false positives as it has very significant illustrative charecteristics. With the implementation of such fast algorithm for text localization and extraction, algorithm can be designed for data-mining the text contents in future using certain content based text extraction approach, it can also be used in real-time monitoring system more effectively with zero error in its results. The various charecteristics of the text contents in the video can be discussed as followings:

(i). Configuration: The configuration will normally consist of size and positioning of the text. The proposed system has considered different types of real-time videos where there is a positive feasibility of occurrence of different sizes of text appearing in the same video frame. The proposed system has also consideration of diversified positioning for making the system more challenging. Majority of the prior research work has not much consideration of positioning. Normally, the textual characters are placed either horizontal or vertical (Japanese text). Certain inclined text can also be observed in most of the TV channel logos. Therefore, in order to make most robust text extraction system, we have considered all the possible orientation of the text for much contrast results. Another consideration for our proposed system is also the distance between two different characters or words, which can be maximum time uniform and sometime non-uniform.
(ii) Speed: This is one of the important consideration where we identify the characters which can exist in different frames of video sometime with or without motion. This phenomenal characteristic will be used for tracking text and its improvement.

(iii) Color: There is also a feasibility that the same text or character may appear in same or different colors in every consecutive video frames.
(iv) Boundary: Majority of the text extraction process is created for non-complex readability which results in tough boundary for textual matter and its complex background too. The proposed system has both of these factors considered.

4. Proposed System

The proposed system highlights a robust technique of text extraction from real time videos using contour detection based algorithm like SUSAN [21]. The proposed system has identified contour points as one of the significant charecteristics for video frame which is associated with maximum changes in luminance or in curve direction [22]. Here the strength of image elements inside the window will be compared with that of the center which is also known as Nucleus. In case the comparison yields smaller value than threshold, then the current picture elements will be considered having the similar strength with the center of picture elements. The USAN (Univalue Segment Assimilating Nucleus) will create region with such elements. The contour detection methodology in the proposed system can be formulated as following algorithm:

Algorithm-1: SUSAN
Objective: The SUSAN edge detector has been implemented using circular windows to give isotropic responses
Input: image, threshold for brightness and USAN
Output: Implementation of SUSAN algorithm by setting up circular mask along with removal of close corners.
Steps:

```
1 Create a function for Susan
2 Check inputs and fill in the blank variables
3       Check if the inputs are empty
4       Convert to double image format
5       Prepare the inputs and outputs
6 Create the brightness look up table (LUT)
7       Set up the variables
8 Set up the circular masks
9 Create 37 pixel circular mask
10 37 pixel circular mask for x
11 37 pixel circular mask for y
12 Compute the USAN response
13      Compute correlation for the window mask w
14      If already too big - ignore it
15 Compute correlation for the window mask wx
16 Compute correlation for the window mask wy
17      Compute the sq response
18              Check the centre of gravity
19      Threshold the response to find the corners
20 Perform nonmaximal suppression (5x5 mask)
21      Find the local maxima
22      Corner must be local maxima
23 Initialized removal of any close corners
24      5x5 neighborhood mask (12 points)
25      Remove close corners
```

The contour function is formulated considering arguments like detected contour point for identification, and width and height of the window. A morphological operation like dilation is performed for getting the associated regions in video frames, once the algorithm successfully accomplishes the contour points with text. In the initial stage of processing, certain non-text area will be removed. Finally the segmentation is created for the multiple lines in the frames using vertical and horizontal ridge of the contour point. Finally, all the identified text regions are resized to initial size of the video frame.

The candidate text area captured is refined using Fuzzy C-Means Clustering algorithm.

Algorithm-2: Fuzzy C-Means Clustering Algorithm
Objective: In fuzzy clustering, each point has a degree of belonging to clusters, as in fuzzy logic, rather than belonging completely to just one cluster. Thus, points on the edge of a cluster, may be in the cluster to a lesser degree than points in the center of cluster.
Input: initialization of clusters.
Output: creation of clusters for text refinement
Steps:

```
1 Create a function for Fuzzy C-Means Cluster
2 Choose a number of clusters.
3 Assign randomly to each point coefficients for being in the clusters.
4 Repeat until the algorithm has converged (that is, the coefficients' change between two iterations is no more than $\varepsilon$, the given sensitivity threshold) :
5 Compute the centroid for each cluster
6 For each point, compute its coefficients of being in the clusters
7 Create cluster
8 Calculate number of rows and columns for subplot
9 Display clusters
```

Algorithm-3: Text Extraction using Fuzzy C-Means Clustering.
Objective: The main objective of this program is to extract text region from the input video and display the sub-images along with implementation of Fuzzy C-Means clustering
Input: video and initialization of threshold for brightness and USAN

Output: The output shows sub-images, with separate letters

Steps:

1 *Input the video file*

2 *Use multimedia reader method to read the video*

3 *Initialize brightness threshold.*

4 *Initialize USAN threshold*

5 *Defined Matrix for non text corner points removal*

6 *Convert the video to frames and read the frames*

7 *Call process frame function*

8 *Use image resize method*

9 *Use sub plot method*

10 *Display the sub-image*

11 *Plot a rectangle*

12 *Region Merging*

13 *Separate colors*

14 *Quantize the colors to 16 x 16 x 16 values*

15 *Plot histogram for R, G and B*

16 *merge all color matrix to form one image*

17 *Extract text region*

18 *Use image crop method*

19 *Apply Fuzzy C-means clustering*

20 *Create cluster*

21 *Calculate number of rows and columns for subplot*

22 *Display clusters*

23 *Show separate letters*

The processing of the algorithm will result in candidate text area. Finally, the clustering of the associated regions are estimated. Finally creation of the sub-images will result in merger image which will also give rise to the quantized image, i.e. the characters are merged in order to generate the text line.

5. Implementation and Results

The framework project work is designed in Matlab in 32 bit system 1.8 GHz with dual core processor where different real time videos are considered for the experiment. The basic graphics video display card of DIAMOND AMD ATI Radeon is used for experimenting on both OS of Windows Vista and Windows 7. The implementation also considers videos with single text, multiple text, text with different sizes of fonts, text with complex and simple background, text with different languages.

For the purpose of the experiment, different types of real time video inputs were considered:

1. Video clips of short / long interval
2. Video clips with fast and slow moving text.
3. Video clips with text appearing in different orientations (horizontal, vertical, slanted).
4. Video clips with multiple languages (English, Hindi)

In the preliminary experiment, a video file in AVI format is chosen from a news channel which was the 1st consideration of the video input. The real time video is read by multimedia reader object. The program initialized the brightness threshold and USAN threshold as 20 and 2000 respectively (in Algorithm 2 and 3). A matrix is created for non-text corner point removal. Various sub-images are created which will all identify the dynamic text appearing at every duration.

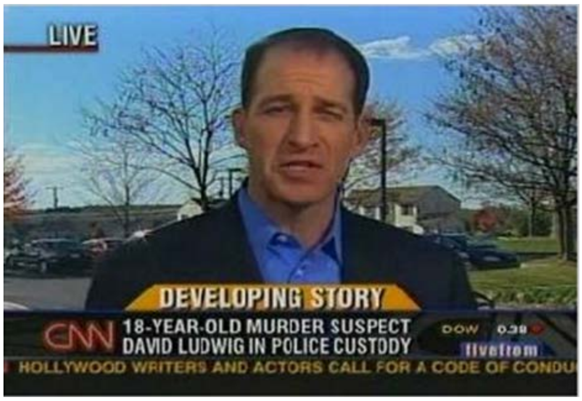

LIVE
DEVELOPING STORY
CNN
18-YEAR-OLD MURDER SUSPECT
DAVID LUDWIG IN POLICE CUSTODY
DOW 0.38
livefrom
HOLLYWOOD WRITERS AND ACTORS CALL FOR A CODE OF CONDU

Fig 1. A video frame.

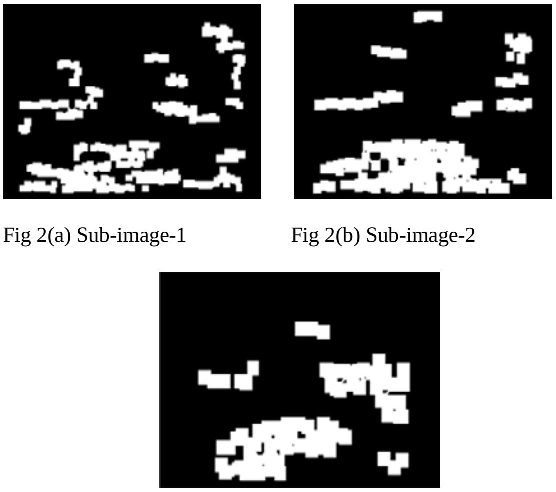


Fig 2(a) Sub-image-1 Fig 2(b) Sub-image-2

Fig 2(c) Sub-image-3

Fig 2. Creation of 3 sub-images from the respective video input as shown in Fig 1.

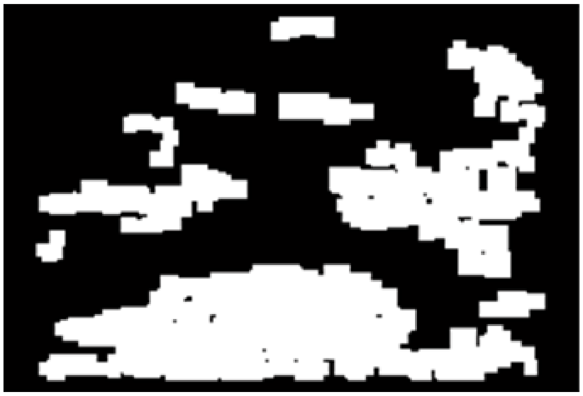


Fig 3. Merger Region output

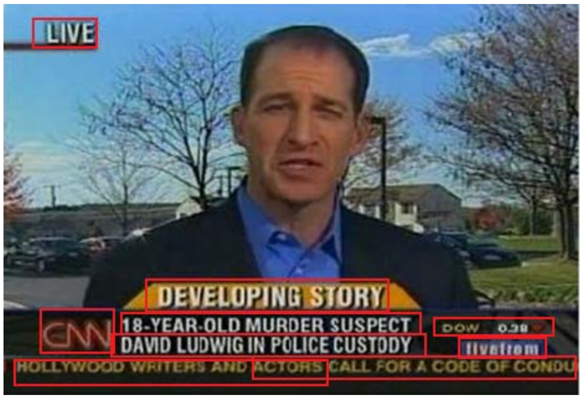

LIVE
DEVELOPING STORY
CNN
18-YEAR-OLD MURDER SUSPECT
DAVID LUDWIG IN POLICE CUSTODY
DOW 0.38
livefrom
HOLLYWOOD WRITERS AND ACTORS CALL FOR A CODE OF CONDU

Fig 4. Text Localization and Extraction

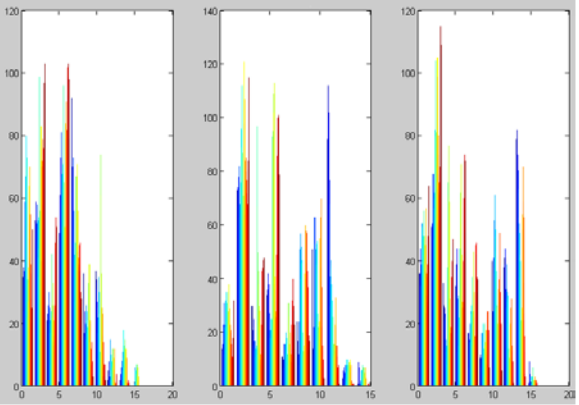


Fig 5. Graph created for Red, Green, and Blue Histogram based color clustering

A text region merging algorithm then identifies the text in each sub-image which will be finally magnified to original size. The R, G, B colors are separated and quantized to 16 x 16 x16 values which gives rise to histogram as shown in Fig 5. Finally all the color matrix are quantized to form one image. As seen in sample news video, that there will exist certain text e.g. Logo text, video caption, time, date, update headings etc, which will not change much for longer duration of play, it might give rise to overlapping and interleaving as the same set of text will be extracted every time by the algorithm. Therefore the intersection regions are selected to avoid this. The final extraction of the text region is done using Fuzzy C-Means Clustering. The segmentation considering all the color constituents is designed with classification based on text and non-text regions. The fuzzy clustering algorithm is one of the efficient data clustering algorithm [23], which is an iteratively most favorable protocol normally with consideration of least square method to design a fuzzy partition of the datasets. The process of iteration halts when the variation between the two consecutive iterations becomes very insignificant. The advantage of this approach is that the two average vectors can be confirmed as the two leading groups e.g. text and non-text regions. Therefore the edge histogram distribution for the detected text regions is estimated, as shown in Fig 5, in order to binarize the text region for effective text extraction.

The second set of the implementation is conducted from the second consideration of video input where the video are sometime very slow and suddenly very fast. This experiment is conducted to measure the efficiency of the algorithm for text extraction for fast moving text in the video. The video input taken for this experiment shows a text “DRENCHING RAIN” and “HOLLYWOOD” which appears onscreen in less than 2 seconds.

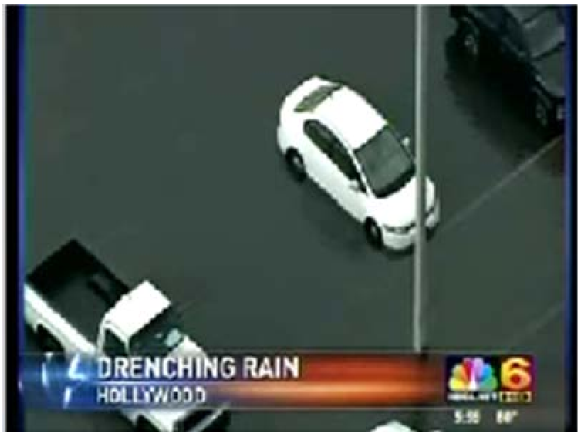


Fig 6. Frame Captured from the video

The application created for this set of evaluation reads the video which calls SUSAN function to identify the corner points using morphological structuring elements. The masking application as shown in Fig 7(b) is created from the corner for each generated frames from the input video. The dilation operation is used here which is used for morphologically open binary image for removing unwanted text region.

Finally, the effective text region is located and extracted as shown in Fig 7(c)-(d).

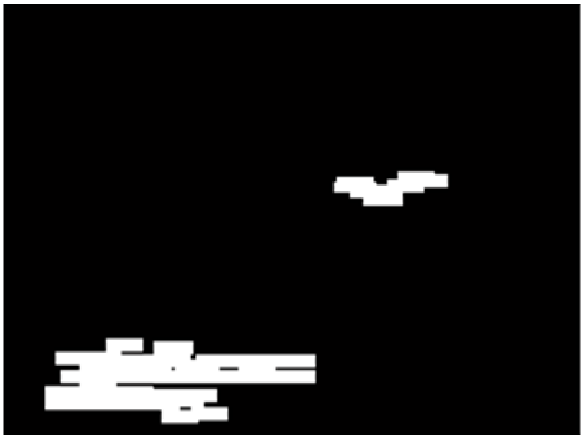


Fig 7(b) Mask Created from corners

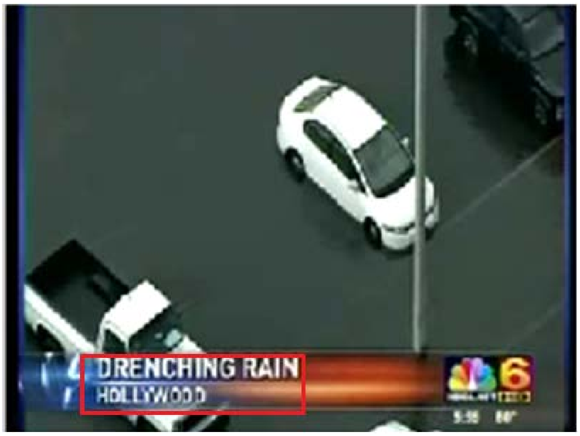


Fig 7(c). Detected Text Region

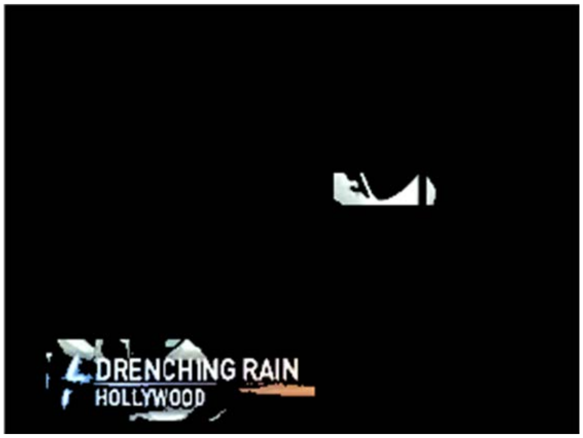


Fig 7(d). Final Output

The next set of the experiment is conducted with 3rd and 4th consideration of the video input. The input video is captured from the TV tuning card for the live broadcasting of the TV show. The significant fact here is the text appearing in this set of video are in different size, style, orientation, as well as language (English and Hindi).

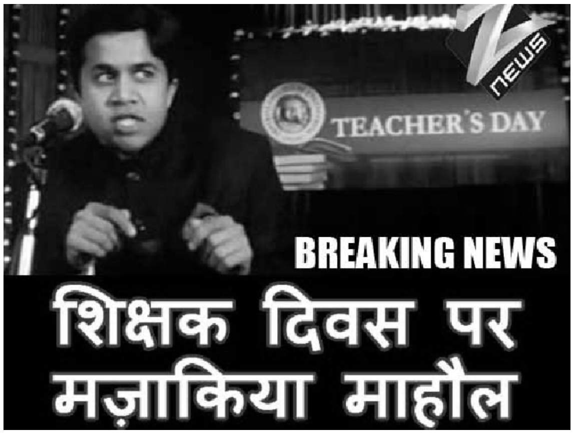


Fig 8 (a). A Video Frame captured

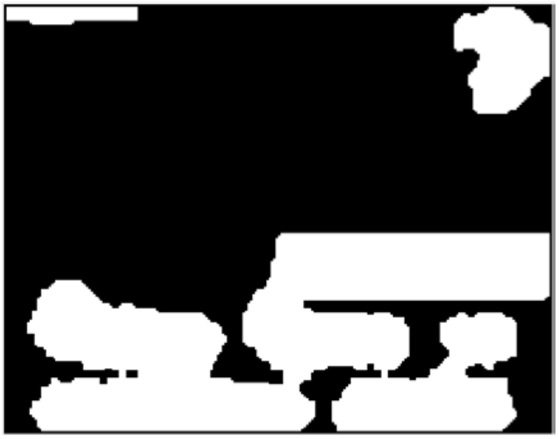


Fig 8(b) Creation of Mask from corners

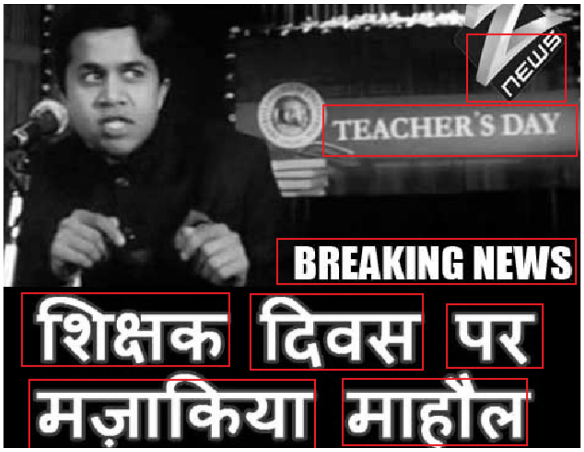


Fig 8 (c). Effective Text Localization

It can be noticed here that almost all the text (both in Hindi and English) is 100% accurately extracted in all implementation of masking. Interesting fact is the English text on the top right side of the channel logo which has different orientation or inclination in comparison to other text is also extract with 95.2 % accuracy. Therefore, the system design for the proposed text extraction can be eventually considered as robust, reliable, and efficient in text extraction in multiple scenarios as understood from the final output as shown in Fig 8(d).

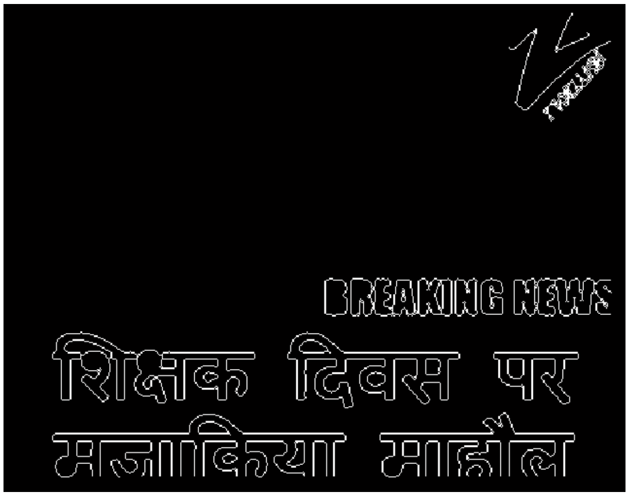


Fig 8 (d). Effective Text Extraction

6. Performance Comparison

For the purpose of comparative analysis, various prior significant researches works in the area of text localization and text extraction has been considered. The proposed system has also being compared and experimented with all the major significant previous research work like:

- *Prior-Work-1*: Qixiang e.t. al (2004) [24] approach has used SVM edge and wavelet features. This is one of the significant work which has better performance in comparison to K-Means clustering. It also has consideration of eastern and western languages.
    - *Comparison to our Approach*: The above discussed approach has no consideration of text orientation as well as fast moving videos. Besides the approach has used multiple number of complex algorithm which consumes much time and definitely not cost effective in terms of text extraction. Our approach has simple use of Morphological operations and Fuzzy C Means Clustering, which effectively works for fast moving videos as well as with text with different orientations.
- *Prior-Work 2*: Matko e.t. al. (2008) [25] approach has worked on player number localization using HSV Color Space and Internal Contours.
    - *Comparison to our Approach*: The result of the above approach has only localization rate (83%) which is considerably higher than recognition rate (52%). This difference is caused by sensitivity of OCR software to non-rigid deformation, noisy character borders etc. Errors in number localization occur has occurred due to skewed numbers, folded jerseys and especially blur which is caused by player or camera motion. But our approach with efficient use of SUSAN algorithm for contour detection has effectively able to overcome all the issues and complexities found in this work.
- *Prior-Work 3*: Shivakumara e.t. al (2009) [26] approach of text detection from video using heuristics rules, initial text block identification, text portion segmentation and new edge features for false positive elimination.
    - *Comparison to our Approach*: The above discussed approach has not considered complex background images. Moreover all the experiment result were derived from testing with English Text and no consideration of text orientation. Whereas our approach has clearly outperformed this results.
- *Prior-Work 4*: Phan e.t. al. (2009) [27] approach of text detection method based on the Laplacian operator and use of K-means Clustering.
    - *Comparison to our Approach*: The work discussed has no consideration of arbitrary orientation. Moreover, the text detection step shows white patches even for non-horizontal text. Where our approach facilitates much more accurate results of text localization and extraction in much more challenging scenarios.
- *Prior-Work 5*: Ghorpade e.t. al (2011) [28] approach with neural network pattern matching technique for text localization, segmentation, and recognition.
    - *Comparison to our Approach*: In this work, as, the characters are recognized on run-time basis, there may be a few cases found in which one or two characters may get misrecognised i.e. a character may get recognized as some other character. Moreover only English text is considered. The process is time consuming due to inclusion of neural network and may not give

the best result in this work. Our proposed work obviously is in much contrast results in comparison to this prior work.

The empirical effectiveness of our proposed algorithm can be derived by estimating the accuracy in detection of the text contents from the input videos in the application designed. The following parameters are considered for this purpose:

- Actually Recognized Text Region (ARTR) consisting of text line.
- False Identified Text Region (FITR) which do not contain any text.
- Missing Text (MT) which ignores some text characters.

The approach manually estimates the Original text region (OTR) where the Text Identification Rate (TIR) can be evaluated as:

TIR = ARTR / OTR

And Error in Text Identification rate (ETIR) as given by,

ETIR = FITR / (ARTR + FITR)

and

Non-Text Identification Rate (NTIR) = MT / ARTR

The comparative analysis is done by estimating the above empirical parameters for the proposed work with all the significant research work specified in this paper. Same sets of the input type and considerations are made towards the analysis.

Table 1. Comparative Performance analysis Parameters with 5 prior research work and proposed system.

Prior Work	**OTR**	**ARTR**	**FITR**	**MT**
Qixiang e.t. al (2004) [24]	500	393	87	79
Matko e.t. al. (2008) [25]	500	349	50	35
Shivakumara e.t. al (2009) [26]	500	251	94	94
Phan e.t. al. (2009) [A27	500	458	39	55
Ghorpade e.t. al (2011) [28]	500	350	55	70
Proposed Work	500	485	38	37

Table 2. Estimation of TIR, ETIR, and NTIR for all the 5 prior research work with proposed system.

Method	**TIR**	**ETIR**	**NTIR**
Qixiang e.t. al (2004) [AR]	78.6	18.12	2.01
Matko e.t. al. (2008) [AR]	69.8	12.53	1.00
Shivakumara e.t. al (2009) [AR]	50.2	27.72	3.74
Phan e.t. al. (2009) [AR]	91.6	7.8	1.2
Ghorpade e.t. al (2011) [AR]	70.6	13.5	2
Proposed Approach	**97.0**	**7.0**	7.0

From the above table, it can be easily identified that proposed system has better Text identification rate (97%) as well as less Error in Text Identification rate (7%) as compared with the prior research work stated.

7. Conclusion

The proposed system highlights a unique technique for text extraction system using contour based SUSAN algorithm and Fuzzy C Means Clustering algorithm. The proposed system is potential enough to extracting text from different types of complex and fast moving text in videos. Even multiple language is also considered for the checking the effectiveness of the algorithm. The proposed system has achieved 97% of Text Identification rate and only 7% of Error in Text Identification Rate. The achieved result is much in contrast compared to majority of prior significant research in this area.

Reference

[1] Dimitris N. Kanellopoulos, Adapting Multimedia Streaming to Changing Network Conditions, IGI Global. 2011
[2] A.K. Jain and B. Yu, “Automatic Text Location in Images and Video Frames”, Pattern Recognition, Vol. 31(12), pp. 2055-2076, 1998.

[3] M. Anthimopoulos, B. Gatos and I. Pratikakis, “A Hybrid System for Text Detection in Video Frames”, The Eighth IAPR Workshop on Document Analysis Systems (DAS2008), Nara, Japan, , pp 286-293, September 2008.

[4] M. R. Lyu, J. Song and M. Cai, “A Comprehensive Method for Multilingual Video Text Detection, Localization, and Extraction”, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 15, No. 2, pp 243-255, February 2005.

[5] C. Liu, C. Wang and R. Dai, “Text Detection in Images Based on Unsupervised Classification of Edge-based Features” , pp. 610-614, ICDAR 2005.

[6] E. K. Wong and M. Chen, “A new robust algorithm for video text extraction”, Pattern Recognition 36, pp. 1397-1406, 2003.

[7] P. Shivakumara, W. Huang and C. L. Tan, “An Efficient Edge based Technique for Text Detection in Video Frames”, The Eighth IAPR Workshop on Document Analysis Systems (DAS2008), Nara, Japan, pp 307-314, September 2008.

[8] Y. Zhong, H. Zhang and A.K. Jain, “Automatic Caption Localization in Compressed Video”, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 22, No. 4, pp. 385-392, 2000.

[9] K. L Kim, K. Jung and J. H. Kim, “Texture-Based Approach for Text Detection in Images using Support Vector Machines and Continuous Adaptive Mean Shift Algorithm”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 25, No. 12, pp 1631-1639, December 2003.

[10] Q. Ye, Q. Huang, W. Gao and D. Zhao, “Fast and robust text detection in images and video frames”, Image and Vision Computing 23, pp. 565-576, 2005.

[11] H. Li, D. Doermann and O. Kia, “Automatic Text Detection and Tracking in Digital Video”, IEEE Transactions on Image Processing, Vol. 9, No. 1, pp 147-156, January 2000.

[12] W. Mao, F. Chung, K. K. M. Lam and W. Siu, “Hybrid Chinese/English Text Detection in Images and Video Frames”, ICPR, Volume 3, pp 1015- 1018, 2002.

[13] H. B. Kekre, V. A. Bharadi, P. Roongta, S. Khandelwal, V. I. Singh, S. Gupta, P. P. Janrao , Performance Comparison of DCT, FFT, WHT, Kekre’s Transform & Gabor Filter Based Feature Vectors for On-Line Signature Recognition, International Conference and Workshop on Emerging Trends in Technology, 2011

[14] T.Pratheeba , Dr.V.Kavitha , S.Raja Rajeswari, Morphology Based Text Detection and Extraction from Complex Video Scene, International Journal of Engineering and Technology Vol.2(3), pp.200-206, 2010

[15] Hiranmay Ghosh, Sunil Kumar Kopparapu, Tanushyam Chattopadhyay, Ashish Khare, Sujal SubhashWattamwar, Amarendra Gorai, and Meghna Pandharipande, Multimodal Indexing ofMultilingual News Video, International Journal of Digital Multimedia Broadcasting Volume 2010.

[16] Sunitha Abburu, Multi Level Semantic Extraction for Cricket Video By Text Processing, International Journal of Engineering Science and Technology Vol. 2(10), 2010

[17] Shwu-Huey Yen, A, Hsiao-Wei Chang, B, Chia-Jen Wang,C, Chun-Wei Wang,d, Robust News Video Text Detection Based on Edges and Line-deletion, WSEAS Transactions on SIGNAL PROCESSING, Issue 4, Volume 6, October 2010

[18] Qiujun Lan, Extraction of News Content for Text Mining Based on Edit Distance, Journal of Computational Information Systems 6:11, pp. 3761-3777, 2010

[19] V.Vijayakumar, R.Nedunchezhian, A Novel Method for Super Imposed Text Extraction in a
Sports Video, International Journal of Computer Applications (0975 – 8887) Volume 15– No.1, February 2011

[20] Stefanos Vrochidis, Ioannis Kompatsiaris, and Ioannis Patras, Utilizing Implicit User Feedback to Improve Interactive Video Retrieval, Advances in Multimedia Volume, 2011.

[21] Shenghua Xu; Litao Han; Lihua Zhang, An Algorithm to Edge Detection Based on SUSAN Filter and Embedded Confidence, Intelligent Systems Design and Applications, ISDA '06. Sixth International Conference, 2006

[22] M. Bertini , C. Colombo , A. Del Bimbo , Via S. Marta, Automatic caption localization in videos using salient points, Proceedings of IEEE International Conference on Multimedia and Expo ICME, 2001

[23] Moh'd Belal Al-Zoubi, Amjad Hudaib, Bashar Al-Shboul, A fast fuzzy clustering algorithm, Proceeding AIKED'07 Proceedings of the 6th Conference on 6th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases - Volume 6, 2007

[24] Qixiang Ye, Wen Gao1, Weiqiang Wang, Wei Zeng, A Robust Text Detection Algorithm in Images and Video Frames, Information, Communications and Signal Processing, the Fourth Pacific Rim Conference on Multimedia. 2003

[25] Matko Šari, Hrvoje Dujmi, Vladan Papi and Nikola Roži, Player Number Localization and Recognition in Soccer Video using HSV Color Space and Internal Contours, World Academy of Science, Engineering and Technology 43, 2008

[26] Palaiahnakote Shivakumara, Trung Quy Phan and Chew Lim Tan, Video Text Detection Based On Filters and Edge Features, Pattern Recognition, 19th International Conference ICPR, 2008.

[27] Trung Quy Phan, Palaiahnakote Shivakumara and Chew Lim Tan, A Laplacian Method for Video Text Detection, 10th International Conference on Document Analysis and Recognition, 2009

[28] Jayshree Ghorpade, Raviraj Palvankar, Ajinkya Patankar and Snehal Rathi , Extracting Text From Video, Signal & Image Processing : An International Journal (SIPIJ) Vol.2, No.2, June 2011

Mr. K N Narasimha Murthy is an experienced teacher for more than two decades and presently working as a Professor and HOD in department of Information Science and Engineering at City Engineering college Bangalore, Affiliated to Visvesvaraya Technological University, India. He got BE degree in 1989, M.Tech degree in 1993, MCA degree in 2003 and M.Tech degree in 2006. His areas of interest are Algorithms, Image processing and Computer Vision and also a member of MIE, MISTE, IACSIT.

Dr. Y S Kumaraswamy working presently as Sr. professor and HOD Department of MCA (VTU) at Dayandasagar college of Engineering Bangalore, India. He has published more than 150 research papers and guided 38 PhD students in the area of Computer Science/ Mathematics and also a selection committee member for ISRO/UGC/DSI profiles

A Novel Method for Efficient Text Extraction from Real Time Images with Diversified Background using Haar Discrete Wavelet Transform and K-Means Clustering

Narasimha Murthy K N[1], Dr. Y S Kumaraswamy[2],
[1] Professor, Dept of Information Science and Engineering, V T U, City Engineering College,
Bangalore, Karnataka, Pin-560062, India,

[2] Professor & HOD, Dept. of MAC (VTU), V T U, Dayananda Sagar College of Engineering,
Bangalore, Karnataka, Pin-560078, India,

Abstract

The proposed system highlights a novel approach of extracting a text from image using two dimensional Haar Discrete Wavelet Transformation and K-Means Clustering. As the commercial usage of digital contents are on rise, the requirement of an efficient and error free indexing text along with text localization and extraction is of high importance. Majority of the previous research work on text extraction has focused on scene text, uniform background, and extensive use of wavelet domain and frequent usage of only grey-scale image as input. The extensive in-depth testing of such approach will lead no not-so-satisfactory results if the image type, non-uniform background, different text orientation, different languages are introduced. The proposed system has broader scale of consideration of input image with much complicated backgrounds along with consideration of sliding windows. For much accuracy, morphological operation is included to accurately distinguish the text and non-text area for better text localization and extraction. The experimental result was compared with all the prior significant work in text extraction where the results show a much robust, efficient, and much accurate text extraction technique.

Keyword: *Text Extraction, Haar, Discrete Wavelet Transform, K-Means Clustering, Morphological Operations*

1. Introduction

Text Extraction from images is a major task in computer vision. Applications of this task are various (automatic image indexing, visual impaired people assistance or optical character reading...). Many studies focus on text detection and localization in images. However, most of them are specific to a constrained context such as automatic localization of postal addresses on envelopes [1], license plate localization [2], text extraction in video sequences [3], automatic forms reading [4] and more generally ”documents” [5]. In spite of such extensive studies, it is still not easy to design a general-purpose TIE system [6]. This is because there are so many possible sources of variation when extracting text from a shaded or textured background, from low-contrast or complex images, or from images having variations in font size, style, color, orientation, and alignment. These variations make the problem of automatic TIE extremely difficult. Increasing popularity of digital cameras and camera phones enables acquisition of image and video materials containing scene text, but these devices also introduce new imaging conditions such as sensor noise, viewing angle, blur, variable illumination etc. Taking into account all these problems and scene text properties it is clear that its extraction and recognition is more difficult task in comparison with caption text and text in documents. Text information extraction consists of 5 steps [7]: detection, localization, tracking, extraction and enhancement, and recognition (OCR). In case of scene text particular focus is set on extraction. This step is done on previously located text area of image and its purpose is segmentation of characters from background that is separation of text pixels from background pixels. Text extraction strongly affects recognition results and thus it is important factor for good performance of the whole process. Text extraction methods are classified as threshold based and grouping-based. First category includes histogram-based thresholding [8], adaptive or local thresholding [9] and entropy-based methods. Second category encompasses clustering-based, region based and learning-based methods. Clustering techniques performed well on color text extraction [10]. Region-based approaches, including region-growing and split and merge algorithm, exploit spatial information to group character pixels more efficiently, but drawback is dependence on parameter values. Learning-based methods mostly refer to multi-layer perceptrons and self-organizing maps, but variation of scene text makes difficult to create representative training database.

The proposed work will introduce novel text extraction techniques with Discrete Wavelet Transform and k-Means Clustering. The system also introduces morphological operation like dilation and erosion for segregation of text and non-text regions for better accuracy. The rest of this paper is organized as follows. We discuss related work in Section II. The research methodology is discussed in Section-III. Proposed system is elaborated in Section IV. Implementation and Results is described in Section-V. Performance Analysis of the proposed system is discussed in Section-VI and finally conclusion and future work is described in Section-VII

2. Related Work

Syed Saqib Bukhari [11] presents a new algorithm for curled textline segmentation which is robust to above mentioned problems at the expense of high execution time. His approach is based on the state-of-the- art image segmentation technique: Active Contour Model (Snake) with the novel idea of several baby snakes and their convergence in a vertical direction only.

Samuel Dambreville [12] has combined the advantages of the unscented Kalman Filter and geometric active contours to propose a novel method for tracking deformable objects. Chen Yang Xu [13] have introduced a new external force model for active contours and deformable surfaces, which we called the gradient vector flow (GVF) field. The field is calculated as a diffusion of the gradient vectors of a graylevel or binary edgemap.

Wumo Pan e.t.al [14] has proposed a novel approach to detect texts from scene images captured by digital cameras. The system converts the text detection issue to a contour classification problem by means of the topographic maps, and performs shape classification by exploiting the over-complete and sparse structure in the shape data.

Fabrizio e.t. al [15] has presented a text localization technique which was considered to be efficient in the difficult context of the urban environment. The system uses a combination of an well-organized segmentation procedure based on morphological operator and a configuration of SVM classifiers with a variety of descriptors to estimate regions that are either text or non-text area. The system is competitive but generates many false positives Baba [16] has proposed a novel approach for text extraction by analyzing the textural evaluation in general scene images. The work has introduced a hypothesis that texts also have equivalent charecteristics that differentiates them from the natural background. The researcher has estimated spatial difference of texture to achieve the distribution of the degree of likelihood of text region.

Aghajari [17] propose an approach to automatically localize horizontally texts appearing in color and complex images. The text localization algorithm achieved a recall of 91.77% and a precision of 96%.

Hrvoje e.t. al [18] propose new method for scene text extraction in HSI color space using modified cylindrical distance as homogeneity criterion in region growing algorithm. The work has also introduced Solution for seed pixel selection based on horizontal projection.

Jayant e.t. al [19] present a novel method for extracting handwritten and printed text zones from noisy document images with mixed content. We use Triple-Adjacent-Segment (TAS) based features which encode local shape characteristics of text in a consistent manner. The experiment was tested with only similar types of text present in page. The system also lags different scripts testing.

Sumit [20] has presented a technique for using soft clustering data mining algorithm to increase the accuracy of biomedical text extraction. The development of the proposed algorithm is of practical significance; however it is challenging to design a unified approach of text extraction that retrieves the relevant text articles more efficiently. The proposed algorithm, using data mining algorithm, seems to extract the text with contextual completeness in overall, individual and collective forms, making it able to significantly enhance the text extraction process from biomedical literature.

3. Research Methodology

The issues of text extraction discussed in this proposed system from given image are multifold and can be segregated for various processing like binarization, implementing wavelet domain, morphological procedures, and finally localization and recognition of text.

The proposed system as shown in Figure 1 presents a research methodology where the text extraction from images with different scenario deploying discrete wavelet transform and k-means clustering. The prominent edges captured from the input binarized image are estimated using two dimensional discrete wavelet transform. Finally, when this stage is accomplished, morphological operations like erosion and dilation is implemented for the purpose of removing some non-text area which can be easily confused as text region. The morphological operations also associated various segregated candidate text regions in each information for sub-band of the binarized image. The fact in this stage for consideration is that binary information about the colors actually do not assist in text extraction procedure from the given image. The proposed system accepts input as colored RGB image for more real-time environment in development. The image is then processed in wavelet domain and then the text extraction process is implemented in later stage of processing. The proposed discrete wavelet transform system can be exhibited by following flow:

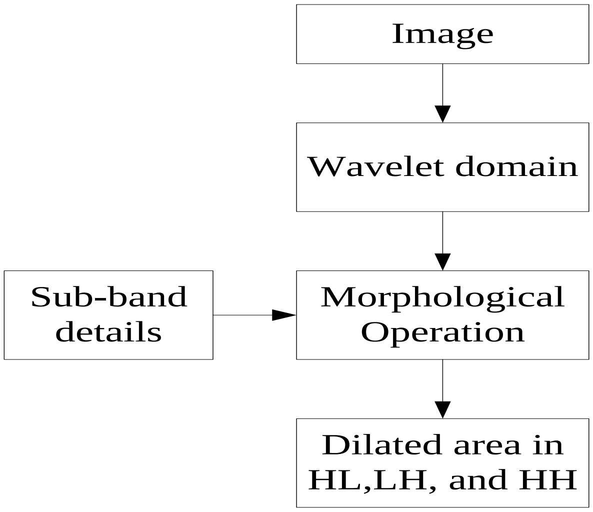


Fig 1. Proposed wavelet based text extraction protocol

A. Discrete Wavelet Transform

Digital image processing has witnessed a discrete wavelet transform as a prime tool in the area of multi-resolution analysis [21]. 1-D discrete wavelet transform decomposes an input image into mean constituent and detail constituent by estimation with the help of high-pass filter and low-pass filter [22]. Whereas 2D discrete wavelet transform will decompose an input image into 4 sub-bands (LL (*mean constituent*), LH, HL, and HH (*detailed constituent*)).

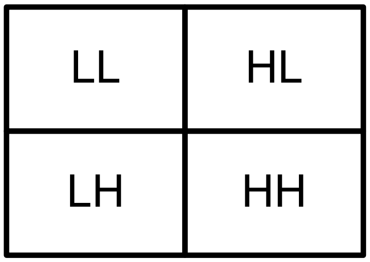


Fig 2. 2-D DWT decomposition output representation

The multi-resolution of the two dimensional wavelet domains can be deployed to explore the text regions of an input image. The conventional filters and detection mechanism for regions can also be expected to provide the equivalent output too. In comparison to one dimensional, 2D discrete wavelet transform can be the better option as it can identify maximum number of edges in one time which cannot be done by conventional algorithms. The conventional boundary detection filters can identify 3 types of boundaries using different types of masking operators as shown in Fig 3. This is also one of the significant reasons of why the conventional boundary detection filters are not faster in comparison to two dimensional discrete wavelet transform.

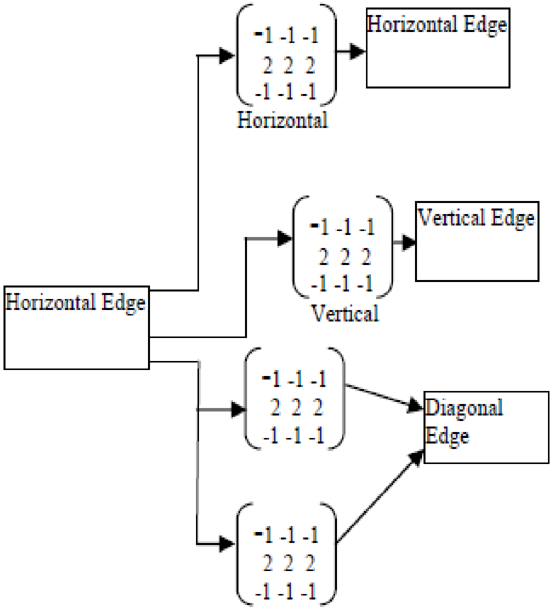
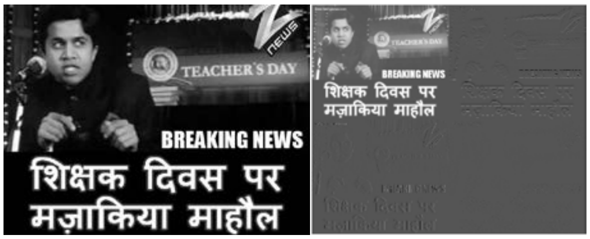


Fig 3. Conventional boundary detection by mask operator



(a) (b)

Fig 4. (a) Actual grey image (b) DWT coefficients

A grey scale image when achieved from the original input of RGB image is as shown in Fig 4(a). Fig 4(b) shows how the discrete wavelet transform converts the gray scale image into four sub-bands. The similar operation when performed by Haar [23] discrete wavelet transform makes the processing less complicated, faster with good accuracy, and efficient in comparison to other types of wavelet domain. The important features of the Haar wavelets are very contributing factors in the proposed methodology. The Haar DWT is genuine, symmetric, and orthogonal with simplest boundary situation along with support for random spatial grid distance. It also supports simple high-pass and low-pass filter coefficient [23].

$$\begin{bmatrix} A & B & C & D \\ E & F & G & H \\ I & J & K & L \\ M & N & O & P \end{bmatrix} \qquad \begin{bmatrix} (A+B) & (C+D) & (A-B) & (C-D) \\ (E+F) & (G+H) & (E-F) & (G-H) \\ (I+J) & (K+L) & (I-J) & (K-L) \\ (M+N) & (O+P) & (M-N) & (O-P) \end{bmatrix}$$

(a) (b)

$$\begin{bmatrix} (A+B)+(E+F) & (C+D)+(G+H) & (A-B)+(E-F) & (C-D)+(G-H) \\ (I+J)+(M+N) & (K+L)+(O+P) & (I-J)+(M-N) & (K-L)+(O-P) \\ (A+B)-(E+F) & (C+D)-(G+H) & (A-B)-(E-F) & (C-D)-(G-H) \\ (I+J)-(M+N) & (K+L)-(O+P) & (I-J)-(M-N) & (K-L)-(O-P) \end{bmatrix}$$

(c)

Fig 5. (a) The source image (b) Row operation in 2-D Haar DWT (c) Column operation in 2-D Haar DWT

A sample of 4x4 grey level images is shown in Fig 5(a). The addition and subtraction is applied on grey scale image for evaluating wavelet coefficient. The two dimensional discrete wavelet transform is accomplished by dual structured one dimensional discrete wavelet transform with both rows and columns. The row operation is conducted first in order to obtain the output as shown in Fig 5(b). Column operation is then used for transformation which finally gives the output of two dimensional Haar discrete wavelet transform as shown in Fig 5(c). A gray-scale image is converted to one mean constituent sub-band and three detail constituent sub-bands using two dimensional Haar DWT. Using Haar discrete wavelet transform on the image, diversified information about the text regions can be identified from the sub-bands details. For an example, LL subband identifies mean constituents, HL sub-bands identifies vertical boundaries, LH sub-bands identifies horizontal boundaries, and HH sub-bands identifies diagonal boundaries. The easy way to understand this is to observe the Fig 4 (a) which is basically a grey-scale image when subjected to Haar discrete wavelet transform gives the output as represented in Fig 5. The candidate text boundaries in the source image can seem from the detailed constituent's sub-bands (HL, LH, and HH).

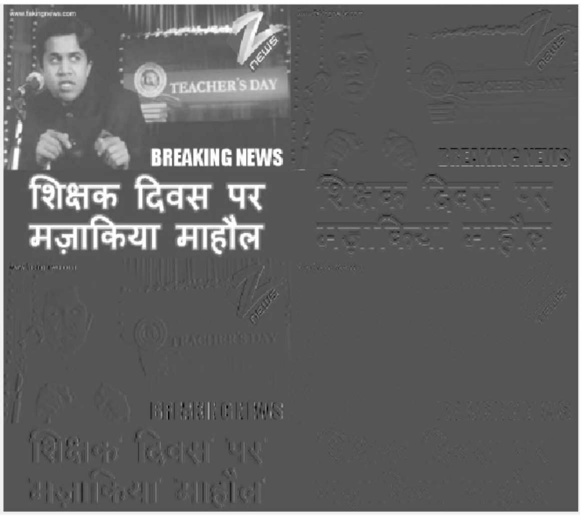


Fig 6. Implementing Haar discrete wavelet transform to source input image

B. K-Means Clustering

The k-means is basically a clustering algorithm which partition a data set into cluster according to some defined distance measure [24][25]. One of the significant tasks in machine learning is to comprehend images and extracting the valuable details. In this direction of analyzing data within the image, segmentation is the first phase to estimate quantity of the object present in an object. K-means clustering algorithm is an unsupervised clustering protocol [25] which categorizes the input data points into multiple types based on their inherent distance from each other. The protocol considers that the data features create a vector space and tries to locate normal clustering in them. The K-means function is given in (1).

$$[\mathit{mu}, \mathit{mask}]=\text{kmeans (ima, k)} \qquad (1)$$

where mu is the vector of class means, mask is the classification image mask, *ima* is the color image and *k* is the number of classes. The points are clustered around centroids in eq. (2) which are obtained by minimizing the objective [25].

Let m = max (*ima*)+1 , then

$$\text{mu} = \{(1:k) * m\} / (k+1) \qquad (2)$$

The maximum function shown above is the maximum value in the in *ima* matrix which represents the colored image in order to achieve the maximum value of the content colors where the color values are revealed as a unit value for all pixel. This stage is done to explicitly describe the maximum number of levels that can be used for estimating the histogram.

The working principle of the k-means clustering algorithm in the proposed system is as discussed below:

i. The histogram of intensities which should highlight estimates of pixels in that specific tone is estimated as shown below

$$n = \sum_{i=1}^{k} m_i \qquad (3)$$

where,

n = total estimates of observations
k = total estimates of tones.

The quantity of the pixels is estimated by the m_i which has equivalent value. The graph created with the help of this is only the alternative way to represents histogram.

ii. The centroid with k arbitrary intensities as in eq. (2) should be initialized.

iii. The following steps are iterated until the cluster labels of the image do not alters anymore.

iv. The points based on distance of their intensities from the centroid intensities are clustered.

v. The new centroid for each of the clusters is evaluated.

C. Morphological Operation

The morphological operations like dilation and erosions are used for better approach of refining text region extraction. The non-text regions are removed using morphological operations. Various types of boundaries like vertical, horizontal, diagonal etc are clubbed together when they are segregated separately in unwanted non-text regions. But, it is also known that the identified region of text consists of all these boundary and region information can be the area where such types of boundaries will be amalgamated. The boundaries with text are normally short and are associated with one other in diversified directions. The proposed system has deployed both dilation and erosion for associating separated candidate text boundaries in every detail constituent sub-band of the binary image.

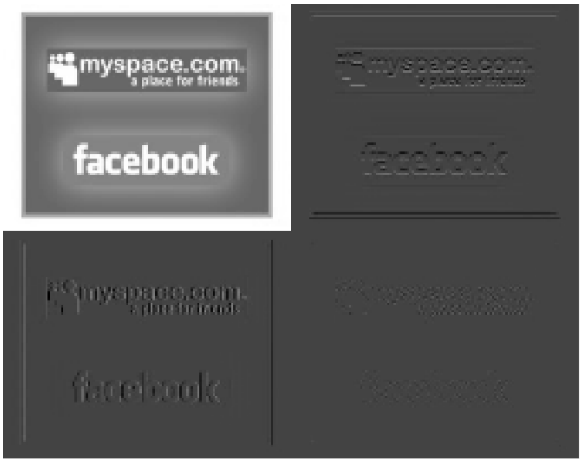


Fig 7 Implementation of Morphological operations on three binary regions

Finally, the morphological operations like dilation and erosion is designed exclusively to fit use-defined input of text based image with various type of charecteristics.

4. Proposed System

The proposed work is designed to accept the input as an image where the final effective output is obtained as extracted text using k-means clustering algorithm and mathematical morphological operations. For contrast in the results, discrete wavelet transform is applied for decomposing the image to sub-bands at various scales with diversified resolution.

The text area is considered as special texture with unbalanced texture charecteristics. Various statistical features like mean, standard deviation, and energy is estimated when the image with text is subjected to discrete wavelet transformation algorithm. After the image is subjected to wavelet transform, classification based on region is applied for compacting the text area within the scope of image. A specific sliding window is designed which reads the high frequency sub bands by sliding steps. The application can be considered that the dimension of each sub-band is M×N after subjecting one-level wavelet transform, and we have,

d_1= mod (M-W, l_1), d_2= mod(N-H, l_2)

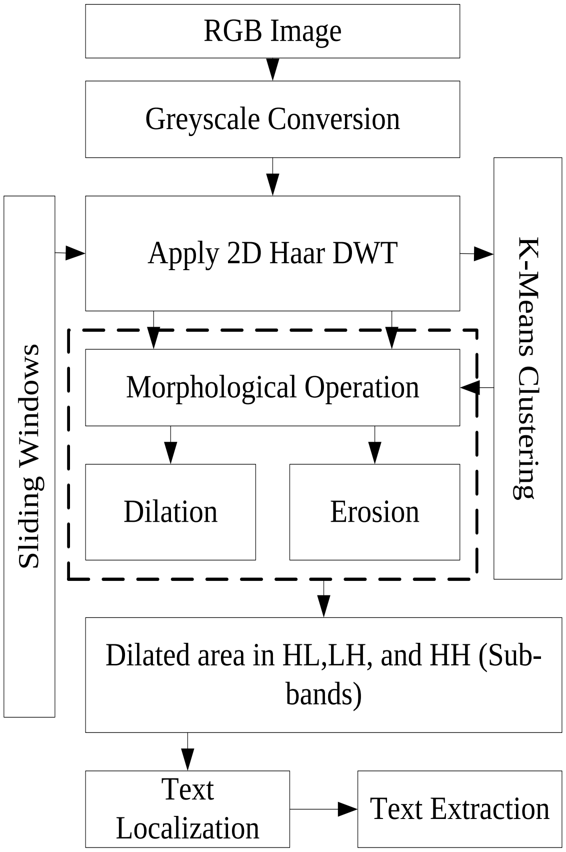


Fig 8. Overall Architecture of the proposed system

if d1 and d2 are not equal to zero, than it fails to superimpose all the area of every sub-band when sliding window reads the high frequency sub-bands by the step l_1 x l_2. The work also rejects all the contents which do not belong to the region.

The statistical charecteristics of every sub-band is estimated. The process achieves 12 features by evaluating the charecteristics of three high frequency subbands. Finally 12- dimension text feature vector is constructed.

The second phase of the design uses k-means clustering protocol where clustering is deployed by analyzing the texture characteristic vector. The clustering factors selected are primary point of text, normal background, and complex background. Care should be taken to update the point of cluster in every processing of k- iterations. The image is segregated into three categories for textual area, simple and complex background area. Binarization technique is applied to the image depending on the results of classification and then mathematical morphological operations are deployed to take out the text details from the image. The effective algorithm implemented in the proposed system is as follows:

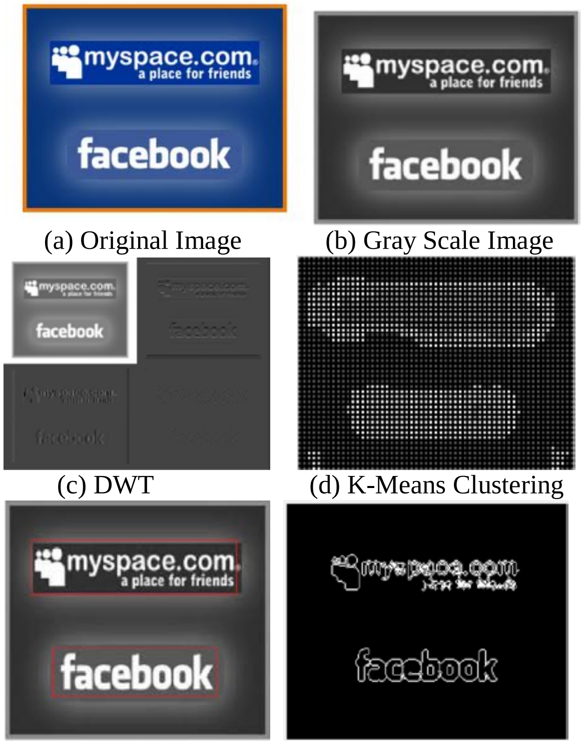
```
START
Input RGB image
If image is RGB
then covert to Gray scale
Create a function for performing DWT
Use Haar 2D DWT
Perform DWT
Initialize the coefficients, sub-bands
Create a function for sliding window
 [W H] =size (window1)
 mu = mean (mean (window1))
 window2 = (window1-mu)
 stanDev= sqrt (sum (sum (window2.^2))/(W*H))
 E = sum (sum (window1.^2))
 Estimate Size of subband
 Create a function for K-Means Clustering
 Calculate column number and row number
 For zero padding
 Apply zero Padding
 Extract the features of sliding window
 Rebuild the cluster id
 Apply Mask Operation
 Morphological operations on binary images
  Detect boundary using Sobel
  Morphologically open binary image
        (remove small objects)
STOP
```

One of the prime issues of implementing clustering algorithm is an inevitable computing error for which reason once the text area is extracted, the system cannot facilitate wholesome error free information about the complete text area. Therefore, the design implements morphological operations like erosion and dilation in order to measure and localize the all text sub-areas. Another issue is the non-text pixels which are also eliminated using erosion and dilation. The appropriate position of the text region is localized in the original image by merging the text pixel locus that is not extracted around the text region boundary. Finally the actual text information is extracted from the processed binarised image.

5. Implementation and results

The framework project work is designed in Matlab in 32 bit system 1.8 GHz with dual core processor where total of 150 different types of images are considered for the experiment. The basic graphics video display card of DIAMOND AMD ATI Radeon is used for experimenting on both OS of Windows Vista and Windows 7. The implementation also considers images with single text, multiple text, text with different sizes of fonts, text with complex and simple background, text with different languages.

The input image binarised to grayscale which is then subjected to discrete wavelet transform. The system then subjects the processed image into k-means clustering protocol. Morphological operation like erosion and dilation is deployed in order to remove all the unwanted non-text region which can be confused with the text regions sometimes. Finally text localization and extraction takes place as shown in the results below:



(a) Original Image (b) Gray Scale Image

(c) DWT (d) K-Means Clustering

(e) Erosion/ Dilation (f) Text Localization

Fig 9. Results from Text Extraction Process

The above results in Fig 9. shows the output of the application obtained when an image of simple background and different multiple text with different font size is used. The localization process along with text extraction is found to be satisfactory.
Our preliminary experiment although was not so satisfactory when the attempt was conducted on the scanned image for text extraction. The accuracy rate was only 75%. The experiment is also conducted in image with text in Hindi language unlike the previous experiment. To identify the robustness and compatibility of the designed application, the experiment was conducted with two set of image e.g.:

- Image with Hindi text with simple background and with same font size.
- Image with both Hindi and English text with different font size and style and orientation.

The second set of the experiment is conducted to scrutiny the efficiency of the protocol towards text extraction for non-English text. Here we chose Hindi language for testing as it is one of the most frequently used language in any type of document related to Indian Government. Fig 10. shows the reliability of the application for extracting the Hindi text. The error percentage is zero in this case showing system to be robust in Hindi Language too along with English. But a fact has to consider that this experiment is conducted with condition of simple background and not all the text will have simple background.

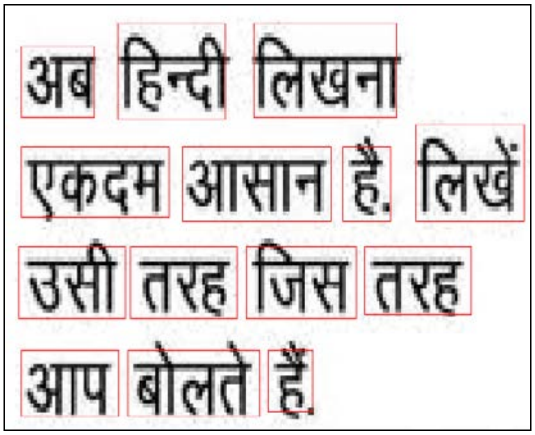
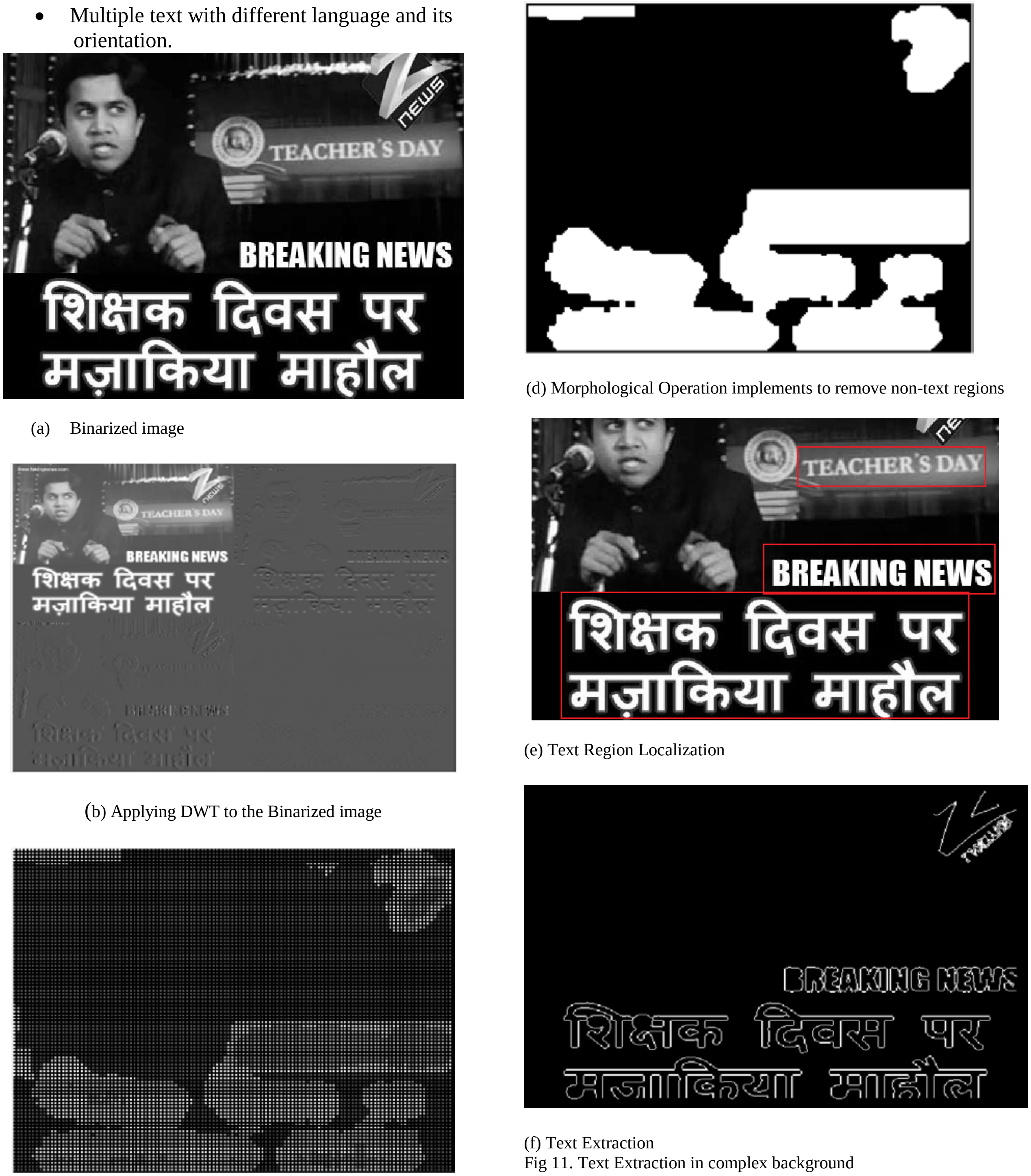


Fig 10. Output of Text Extraction for Hindi Font with simple background

The third set of the performance analysis is conducted considering complex background. Complex background can be defined as an image with high variation of RGB along with illumination factor in its background whereas in simple background it is uniform. Therefore, it was a bit of challenging task to have proper consideration of image with multiple text of different font as well as with complex background. So for this set of experiment, we have selected an image captured from the running live video streaming from using TV tuning card. A good graphics adapter will be required for proper restoration of the captured image.
The image for this set of experiment is considered as an image with:

- Multiple Text
- Multiple Text with different font size

- Multiple text with different language and its orientation.



(a) Binarized image

(b) Applying DWT to the Binarized image

(c) After implementing k-Means Clustering Protocol

(d) Morphological Operation implements to remove non-text regions

(e) Text Region Localization

(f) Text Extraction
Fig 11. Text Extraction in complex background

It can be noticed here that almost all the text (both in Hindi and English) is 100% accurately extracted in all configurations of the sliding window. Interesting fact is the English text on the top right side of the channel logo which has different orientation or inclination in comparison to other text is also extract with 95.2 % accuracy. Therefore, the system design for the proposed text extraction can be eventually considered as robust, reliable, and efficient in text extraction in multiple scenarios.

6. Performance Analysis

For evaluation purpose, the proposed system deploys actual rectified size and error detected size to analyze the simulation results accomplished.

Actual Rectified Size = $(A_R / A_T) \times 100\%$
Error Detected Size = $(N_T / (A_R + N_T)) \times 100\%$

Where
A_R = Actual Extracted Text Region Size
N_T = Non-Text Region Size
A_T = Actual Text Region Size

The analysis results are optimum when actual rectified size is greater than error detected size. The experiment when evaluated with total of 150 images gives the following results as depicted in Table 1.

Table 1: Analysis Results

H	W	Error Detected Size	Actual Rectified Size
16	32	13.6	94.5
16	16	8.5	91.8
8	16	5.1	88.6
8	8	4.8	86.3

The above results interpret that using the proposed algorithm assist the applications to achieve reduction in error detection size and higher actual rectified size. The process successfully extracts the text from various sets of experimental images with the sliding window size of smaller dimension.

Comparison with Techniques
Majority of the research work conducted in past has used datasets of images. But the proposed work is totally focused on real time images being captured from digital camera, or mobile phone, or from any image capturing devices for better study for realistic result. The current research work has been compared with certain conventional algorithm for cross-checking its efficiency. All the analysis process employs the exploration of accuracy of text and non-text area for the given colored image input. The method for checking performance has deployed the complex methodology of discovering all the prominent boundaries and contours at different orientations considering images with multiple text with different font size, style, and language (English, Hindi) for the proposed process of text extraction. It has been noticed that the complexity of the applied protocol increases in order to recognise the boundaries at multiple different directions.

The prominent morphological operator like dilation is employed for sequencing the clusters of the segregated text to a significant complete word or sentences. The proposed system also explores successfully the quantity of the constituents and estimates the degree of inconsistency for each constituent variance. The consideration of detection of text is symbol in case the value of the inconsistency of each constituent is greater. The proposed algorithm though it was found to be very sensitive to skew and direction of placement of text, but the result accomplished in majority of test on 150 images were found to be successful. One of the prime intentions behind the proposed text extraction system is to diminish the probability of detection of non-text elements from the test image. The effectiveness of the morphological operations was also tested by analyzing their respective output image.

The proposed system has also being compared and experimented with all the major significant previous research work like:

- Morphological approach considering scene text by Hasan and Karam (2000) [26]
- Wavelet based feature extraction and neural network for texture analysis considering slanted scene text, localization, and tracking by Li et. al (2000) [27]
- Text detection and localization using DCT coefficient and macroblock type information by Lim et. al (2000) [28]
- Gabor filter like multi-layer perceptron for texture analysis by Jung e.t. al (2001) [29]
- Text detection using sparse representation by Pan et. al (2009) [14]
- Text Localization algorithm in color image via New projection profile by Aghajari (2010) [17]
- Text extraction using data mining algorithm by Sumit (2011) [20]

All the above research work done has extensively used the thresholding concepts along with morphological operations. Whereas the proposed system has contrast implementation of the above mentioned work along with novel introduction of k-means clustering and Haar discrete wavelet transform of two dimensional. All the 150 images has been tested with the above mentioned research work and compared with the proposed work to observe that majority of the experiments with the previous approach when used, it gives better results only in case of consistent background. One more prime observation is that when the images with multiple text style, size, orientation and especially languages are used, all the previous approaches yields false results of text extraction. For example, when the above mentioned research works is implemented on the same test image, the results obtained are as followings:

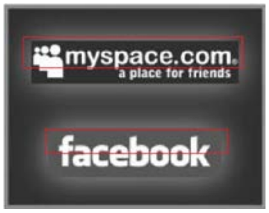


Fig 12(a)

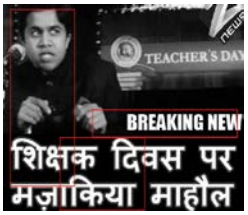


Fig 12(b)

The issue observed above is for the image with uniform background (Fig 12(a)) the text localization is somewhat valid, but it fails to locate the text when background is making inconsistent like in Fig 12(b). There is much such type of errors in the results obtained when previous approaches are compared with the existing one. The proposed protocol for text extraction is therefore considered to be robust and efficient.

7. Conclusion

The proposed system has introduced a novel process of text extraction considering multiple cases of image with its textual contents. The system has been implemented using 2D Haar DWT along with k-means clustering algorithm. It also deploys methodology of sliding window for reading sub-bands of high frequency. Morphological operations like dilation and erosion has been introduced finally to refine the text and non-text region appropriately. For more realistic and robust results, the proposed system has been experimented with images with single / multiple text, multiple text of different sizes / style / languages, images with uniform and non-uniform background. The system is also evaluated with major research results in the past for conventional text extraction approach and is found to be potential for more accurately extracting text information. The future work will be to extending the similar concept of extracting text from video with higher accuracy.

Reference

[1] Palumbo, P. W., Srihari, S. N., Soh, J., Sridhar, R. and Dem-janenko, V., 1992. Postal address blocks location in real time. Computer 25(7), pp. 34–42.

[2] Arth, C., Limberger, F. and Bischof, H., 2007. Real-time license plate recognition on an embedded DSP-platform. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR '07) pp. 1–8.

[3] Wolf, C., Michel Jolion, J. and Chassaing, F., 2002. Text localization, enhancement and binarization in multimedia documents. In: In Proceedings of the International Conference on Pattern Recognition (ICPR) 2002, pp. 1037–1040.

[4] Kavallieratou, E., Balcan, D., Popa, M. and Fakotakis, N., 2001. Handwritten text localization in skewed documents. In: International Conference on Image Processing, pp. I: 1102–1105.

[5] Wahl, F., Wong, K. and Casey, R., 1982. Block segmentation and text extraction in mixed text/image documents. Computer Graphics and Image Processing 20(4), pp. 375–390.

[6] Keechul Junga, Kwang in Kimb, Anil K. Jainc, Text information extraction in images andvid eo: a survey, Pattern Recognition 37 (2004) 977 – 997

[7] H.J. Zhang, Y. Gong, S.W. Smoliar, S.Y. Tan, Automatic parsing of news video, Proceedings of IEEE Conference on Multimedia Computing and Systems, Boston, 1994, pp. 45–54.

[8] A.W.M. Smeulders, S. Santini, A. Gupta, R. Jain, Content-basedimage retrieval at the endof the early years, IEEE Trans. Pattern Anal. Mach. Intell. 22 (12) (2000) 1349–1380.

[9] M.A. Smith, T. Kanade, Video skimming for quick browsing basedon audio andimage characterization, Technical Report CMU-CS-95-186, Carnegie Mellon University, July 1995.

[10] M.H. Yang, D.J. Kriegman, N. Ahuja, Detecting faces in images: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2002) 34–58.

[11] Syed Saqib Bukhari, Coupled Snakelet Model for Curled Textline Segmentation of Camera-Captured Document Images, Proceedings of the 2009 10th International Conference on Document Analysis and Recognition, 2009

[12] Samuel Dambreville, Tracking deformable objects with unscented Kalman filtering and geometric active contours, American Control Conference, IEEE, 2006

[13] Xu. C, Prince. J, "Snakes, shapes, and gradient vector flow" , IEEE Transaction on Images Processing, Vo!. 7, 1998, pp. 359-369

[14] Wumo Pan, T.D. Bui, C.Y. Suen, Text detection from natural scene images using topographic maps and sparse representations, IEEE, 2009

[15] J. Fabrizio, M. Cord, B. Marcotegui. Text Extraction from Street Level Images, CMRT09 - CityModels, Roads and Traffic 2009. Paris, France.

[16] Baba, Y.[Yoichiro], Hirose, A.[Akira], Spectral Fluctuation Method: A Texture-Based Method to

Extract Text Regions in General Scene Images, IEICE(E92-D), No. 9, September 2009, pp. 1702-1715
[17] G. Aghajari, J. Shanbehzadeh, and A. Sarrafzadeh, A Text Localization Algorithm in Color Image via New Projection Profile, Proceedings of The International Multi conference of Engineers and Computer Scientist, Vol-2, 2010
[18] Hrvoje Dujmić, Matko Šarić, Joško Radić, Scene text extraction using modified cylindrical distance, Proceeding NNECFSIC'12 Proceedings of the 12th WSEAS international conference on Neural networks, fuzzy systems, evolutionary computing & automation, World Scientific and Engineering Academy and Society (WSEAS), 2011
[19] Jayant Kumar; Rohit Prasad; Huiagu Cao; Wael Abd-Almageed; David Doermann; Premkumar Natarajan, Shape codebook based handwritten and machine printed text zone extraction, Proceedings Vol. 7874, Document Recognition and Retrieval XVIII, 2011
[20] Sumit Vashishta, Yogendra Kumar Jain, Efficient Retrieval of Text for Biomedical Domain using Data Mining Algorithm, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 4, 2011
[21] R. Pradip Kumar, P, Nagabhushan, Multiresolution Knowledge Mining using wavelet transform, Engineering Letter, 2007
[22] Abhayaratne, G.C.K., Discrete wavelet transforms that have an adaptive low pass filter, Signal Processing and Its Applications, 2003. Proceedings. Seventh International Symposium on
Issue Date: 1-4 July, Vol 2, 2003
[23] Patrick J. Van Fleet, Discrete Haar Wavelet Transforms, PREP - Wavelet Workshop, 2006

[24] S K Gupta, K Sambasiva Rao and Vasudha Bhatnagar, K-means Clustering Algorithm for Categorical Attributes, DataWarehousing and Knowledge Discovery Lecture Notes in Computer Science, 1999
[25] Hassana Grema Kaganami, Zou Beiji, M Sami Soliman, Optimal Color Image Enhancement Using Wavelet and K-means Clustering, International Journal of Digital Content Technology and its Applications. Volume 5, Number 1, January 2011
[26] Y.M.Y. Hasan, L.J. Karam, Morphological text extraction from images, IEEE Trans. Image Process. 9 (11) (2000) 1978–1983.
[27] H. Li, D. Doerman, O. Kia, Automatic text detection and tracking in digital video, IEEE Trans. Image Process. 9 (1) (2000) 147–156.
[28] Y.K. Lim, S.H. Choi, S.W. Lee, Text extraction in MPEG compressed video for content-based indexing, Proceedings of International Conference on Pattern Recognition, 2000, pp. 409–412.
[29] K. Jung, Neural network-basedtext location in color images, Pattern Recognition Lett. 22 (14) (2001) 1503–1515.

Mr. K N Narasimha Murthy is an experienced teacher for more than two decades and presently working as a Professor and HOD in department of Information Science and Engineering at City Engineering college Bangalore, Affiliated to Visvesvaraya Technological University, India. He got BE degree in 1989, M.Tech degree in 1993, MCA degree in 2003 and M.Tech degree in 2006. His areas of interest are Algorithms, Image processing and Computer Vision and also a member of MIE, MISTE, IACSIT.

Dr.Y S Kumara swamy working presently as Sr. professor and HOD, Department of MCA (VTU) at Dayandasagar college of Engineering Bangalore, India. He has published more than 150 research papers and guided 38 PhD students in the area of Computer Science/ Mathematics and also a selection committee member for ISRO/UGC/DSI profiles

Configuration of FPGA for Computerized Speech/Sound Processing for Bio-Computing Systems

V. Hanuman Kumar and P. Seetha Ramaiah

Department of Computer Science & Systems Engineering,

Andhra University, Visakhapatnam , India, 530003

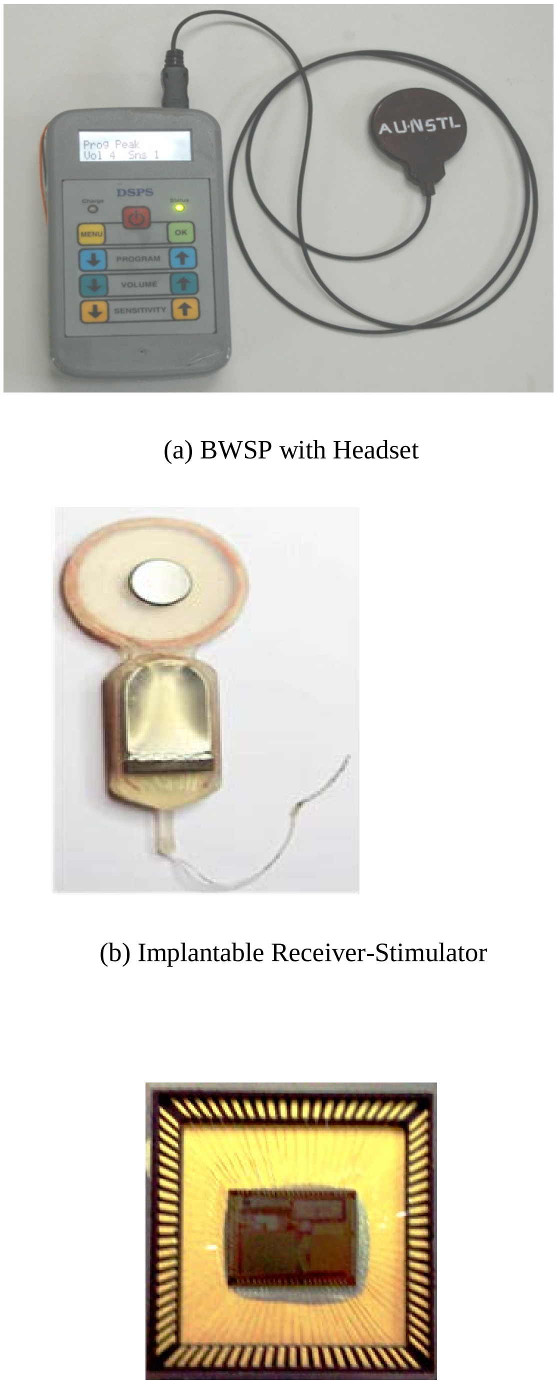
Abstract

The development of Embedded Computer based bio-computing systems mimicking the natural functionality of human parts, is in continuous research because of advent of technology that used Very Large Scale Integration (VLSI) devices such as Field Programmable Gate Array (FPGA) to meet the challenging requirement of providing 100% functionality of the damaged human parts. The evolution of Field Programmable Gateway Array (FPGA) devices to the current state- of-art System-On-Chip (SOC) devices poses considerable problems in terms of ensuring designer productivity in developing high end Computerized Biomedical Speech Processing applications to these devices. Modern Programmable FPGA structures are equipped with specialized Digital Speech Processing embedded blocks that allow implementing digital speech processing algorithms with use of the methodology known from digital signal processor these programmable FPGA architectures give the designer the possibility to increase efficiency of the designed system. This paper presents the details of design and development of one of the bio-computing systems such as Bionic Ear or Cochlear Implant **(CI)** system with greater emphasis on configuration of FPGA with efficient processing algorithm. Bio-computing system incorporates Xilinx Spartan3 FPGA as the main chip for DSP IP cores, 32k words of memory, a 16-bit Analog to Digital converter fixed gain amplifier and programmable gain amplifier and transmitter which convey control codes to the receiver stimulator of the cochlear implant .The processor is battery powered and has been programmed to emulate the continuous interleaved sampling speech processor of 8 electrode implant. Here the Bio-computing system is a Real-Time Embedded Computing System **(RTECS)** that is required to collect the real-time speech/sound data, process the data by using speech/sound processing algorithm(s) and send the processed speech data to the electrode array inserted in the damaged inner ear (cochlea) for providing the speech recognition to the deafened person via inductive transcutaneous RF link. This process should run continuously without loss of speech/sound information.

Key words: FPGA, DSP IP, Bio-computing system, Speech Processing algorithm, FPGA VLSI, System-on-chip, RTECS, CI

1. INTRODUCTION

The Bio-computing system comprises of external Body Worn Speech Processor **(BWSP)**, external Impedance Telemetry **(IMT)** and internal Implantable Receiver Stimulator **(IRS)** with an electrode array. BWSP receives an external sound or speech and generates encoded speech data bits for transmission to IRS via radio frequency transcutaneous link for exciting the electrode array by continuously executing speech/sound processing program embedded in BWSP. The IRS receives the ASK modulated encoded speech/sound information, demodulates the ASK signal, decodes the information and stimulates the selected electrode in the electrode array with the appropriate electric stimuli as bi-phasic current pulses by continuously executing decoded speech/sound and electrode stimulation program embedded in IRS to recognize the speech/sound by the deafened person. External IMT is used to measure electrode-tissue impedances of the inserted electrode array inside the damaged cochlea for configuring the number of active electrodes. Commercially available devices are often found to provide little of the flexibility required for use in a research environment, so the need for a fully configurable FPGA based computerized speech/sound processor for use with bio-computing devices is evident. Previously developed portable digital sound processor, referred to as the P-DSP[1], and a modification known as the P-DSP/HA[2], are much larger, heavier and expend more batteries than current commercially available processors. These factors are inconvenient to the hearing-impaired user, and thus may reduce the amount of experience gained by using the processor in everyday conditions away from the laboratory.



(a) BWSP with Headset

(b) Implantable Receiver-Stimulator

(c) Preliminary Version of ASIC Chip for Receiver Stimulator

Figure1: Prototype to Product Development of AU-NSTL Cochlear Implant System

This paper addresses the Design and development of Bio-computing system as a laboratory model based on Xilinx Spartan3 FPGA (1.2V core) as the main chip for DSP IP cores and Micro blaze / Pico blaze microprocessor. The hardware as well as software design and performance issues are also covered in the present paper. The Bio-computing system comprises of the following hardware modules with relevant embedded software control: a) Configurable FPGA for Speech/Sound Processing as BWSP, b) Implantable Receiver-Stimulator and c) Impedance Telemetry.

A configurable FPGA for speech processing consists of a microphone, fixed gain amplifier, programmable Gain amplifier,16-bit ADC, Xilinx Spartan3 FPGA (1.2V core), radio frequency transmitter ,laboratory model of receiver-stimulator and simulated electrode array with a high speed data acquisition system. Figure 1 shows our proposed prototype to product model of Cochlear Implant System designed and developed by AU-NSTL team. The BWSP comprises Ana log Front End (AFE), Digital Speech/Sound Processing system, Speech Data Encoder, and a Radio Frequency Transmitter using Amplitude Shift Keying (ASK) modulation. FPGA can be configured to implement 4-channel to 8-channel **Continuous Interleaved Sampling** (CIS) algorithm based on the patient's active electrodes. This BWSP generates continuous serial TTL data bits based on the voiced or unvoiced signals served as modulating signal to the ASK modulator at 4MHz RF carrier. The ASK signal is applied to the RF transcutaneous link that in turn used to stimulate the Cochlear Implants Implantable Receiver-Stimulator fabricated as laboratory prototype model for testing and validation. The laboratory model of Receiver-Stimulator consists of radio frequency ASK receiver, speech data decoder, stimulus buffer, 8-bit Digital to Ana log Converter(DAC), constant current generator, active electrode selection logic for switch matrix driver, 8-bit Ana log to Digital Converter(ADC), switch matrix based on H-bridge architecture and 12 simulated electrode resistance array. The development of Bio-computing system as per the design has met the requirements of processing real-time speech/sound signals using the principles of embedded computing architecture. The testing and validation of the developed prototype of Bio-computing system is done by start small approach that tests individual functional units followed by an integrated testing. The performance of the developed system is compared with commercially available CI systems and found equivalent performance using relevant test data, simulation tools and emulation tools. The results of experiments with simulated speech/sound test data and real time speech /sound data are enumerated. Finally, the concluding remarks and future directions for an advanced Bio-computing system are addressed.

2. HARDWARE DESIGN

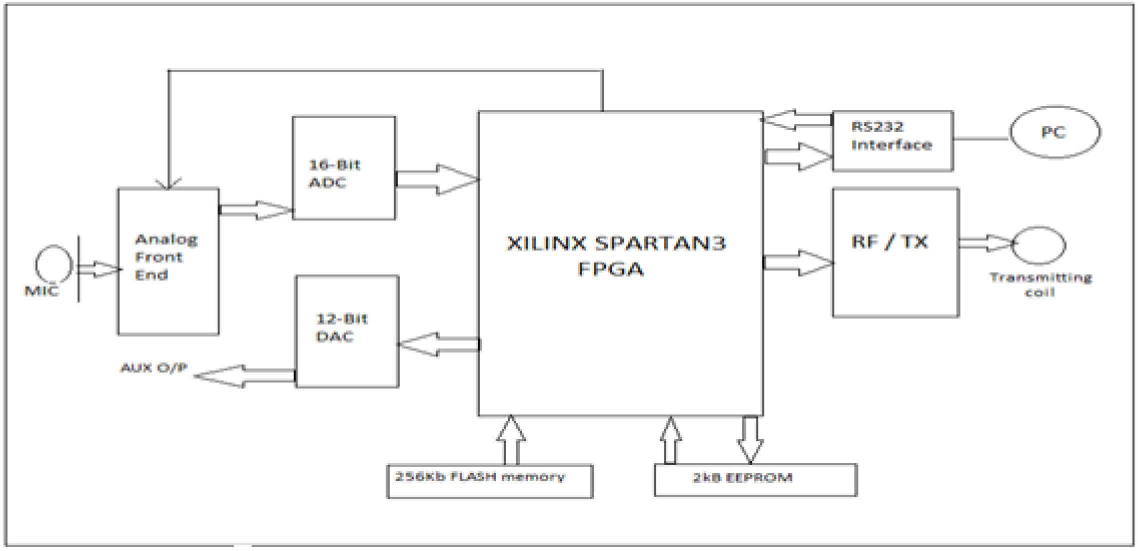


Figure 2: Functional Block Diagram of Configurable FPGA for Bio-computing systems for Speech/sound processing.

The functional components used to configure FPGA for Bio-computing systems can be observed in fig2.Configuration of FPGA for Bio-computing systems to perform speech/sound processing can accept speech or audio signals and transform them into human understandable processed speech or audio to an implantable Bio-computing system's receiver-stimulator of 12 electrodes for making the deaf person to understand the speech or audio tones is designed. The main principle behind the configuration of FPGA involves capturing sound from the environment, processing sound into digital signals and delivering sound to the hearing nerve via electrode array in cochlea. The speech processing system can drive a hearing aid receiver stimulator and excite 8-channel electrode array. In a typical application the system works in the following way.

Sound waves are converted to electrical signals by the microphone and then fed to Ana log Front-end circuit. An electric condenser microphone can be connected to the front panel auxiliary input socket. The sound signals are amplified by fixed gain amplifier with a fixed gain of 30dB and based on volume control of speech processing device, programmable gain amplifier amplifies the output of the fixed gain amplifier and then the signal is attenuated by the sensitivity potentiometer. The signal is filtered to eliminate noise before being converted to a 16-bit digital value by the 16-bit ADC; the 16-bit sample is transmitted to the Xilinx spartan3 FPGA device via a serial interface. The Xilinx spartan3 FPGA typically stores this sample in memory for future processing and may transfer a modified sample back to the SCI to be transmitted to the DAC, where the sample is converted to an Ana log signal to drive a hearing aid receiver. For auditory prostheses use, the Programmable Xilinx spartan3 FPGA based processor periodically construct data frames in the special format required for the cochlear implant receiver

Figure 3: Simplified Block Diagram of the Bio-computing System

stimulator which determines the active electrodes and their current levels and then sends the encoded data frames serially with 171 Kbps rate to the RF transmitter.

The RF transmitter is based on ASK modulation and operates at 4MHz carrier frequency. The RF transmitter modulates the incoming encoded serial data. The encoded data would send to RF transmitting coil. The RF transmitting coil is seventeen turns, 175 strands Litz wire with High Q value. RF transmitter sends the data frames via the transcutaneous inductive coupling to the receiver coil in the laboratory model receiver-stimulator. The receiver stimulator decodes the data and activates the specified electrodes, which stimulate nearby auditory neurons, giving the user the sensation of hearing. A simplified hardware functional block diagram of the Bio-computing system is shown in Figure 3.The Xilinx spartan3 FPGA is used as the central processing system running at a rate of 326 MHz of core clock frequency and Densities as high as 74,880 logic cells. Up to 1872 Kbits of total block RAM ,up to 520Kbits of distributed RAM, Three separate power supplies for the core (1.2V), IOs (1.2V to 3.3V), and Special function(2.5V) eliminating the need for power- consuming external RAM in auditory implant applications. The on-chip peripherals consist of SCI-a Serial Communications Interface, a parallel Host Interface, and a Timer Module and relevant control signals are used to access external memories, as well as the encoder in this application. The required software programs like CIS, ENCODING module stored in an external PROM and these are used by FPGA for speech processing. The mode pin logic levels are automatically altered when the programming cable is connected to the processor. The speech processing Software CIS for DSPMAP is developed in personal computer (PC) using verilog and configured Xilinx spartan3 FPGA using JTAG cable using Xilinx ISE Environment.

3. SOFTWARE DESIGN

There is a huge demand for low cost and high performance of Bio-computing systems in developing countries. Several researchers proposed and attempt to develop a low-cost cochlear implant system but none of these products are available in the market. The development of Bio-computing system involves the strategies of mechanical engineering, physiology, electronics engineering and computer science and engineering. Implementation of embedded speech processing algorithms plays an important role in the development of different techniques for deriving electrical stimuli from the speech signal. Developing speech or sound signal processing algorithms that would help in mimicking the function of a normal cochlea in inner ear is the biggest challenge for the computer engineers. Popular speech processing algorithms such as SMSP, SPEAK, CIS and ACE are used by various vendors are described by the several developers with less technical , design and implementation details due to the limitations of proprietary information or intellectual property rights. The functional block diagram is shown in Figure 4 contains the BWSP, IRS and Electrode array. The Implantable Receiver Stimulator Software decodes the encoded speech signal, generates the required control signals for selection of active electrode for stimulation according to the intensity of the incoming speech/sound signal by means of charge balanced bi-phasic pulses for understanding the speech. The functional block diagram of the Bio-computing system is shown in Figure 5.

Figure 4: Functional Block Diagram of the Cochlear Implant System

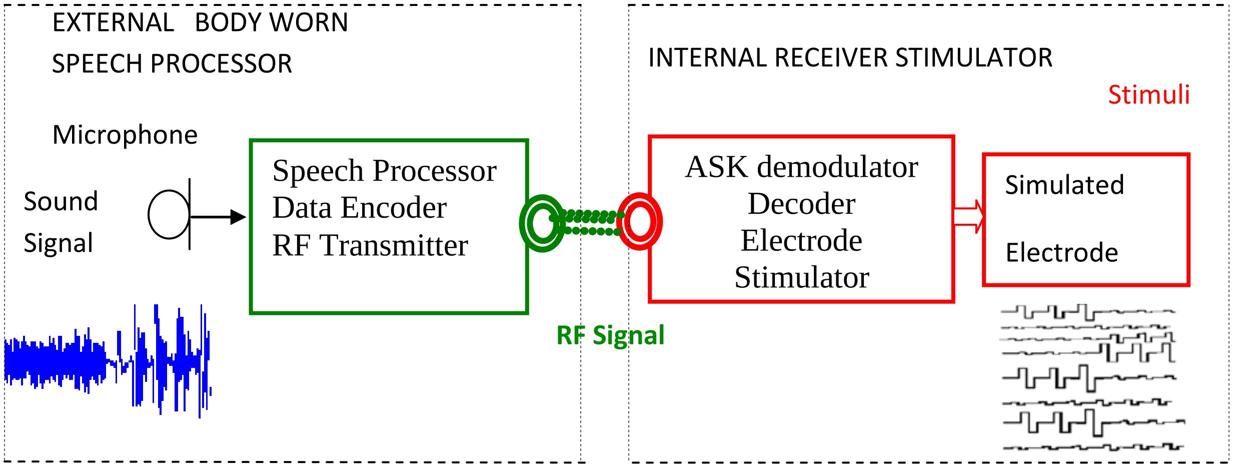


Figure 5: Simplified Functional Block Diagram of Bio-computing

The main functional requirement of the system is to receives the speech/sound signals from the environment and stimulate the electrodes inserted in the cochlea by charge balanced bi-phasic electrical signals. Body Worn Speech Processor receives the electrical signal from the microphone, sampled by CODEC, processed by implementing Continuous interleaved sampling (CIS) speech processing algorithm in FPGA Structures , encoded in using a simple protocol and transmitted serially at 172Kbps rate to the ASK modulated RF transmitter.

Block diagram for 8 –channel CIS implemented in ADSP 2185 is shown in Figure 6. Incoming speech signal is sampled by the ADC is passed to 8 band pass filters. The envelopes of the filtered waveforms are then extracted by full-wave rectification and low-pass filtering (typically with 200 or 400 Hz cutoff frequency). The envelope outputs are finally compressed using nonlinear compression function (e.g., logarithmic) to ensure that the envelope outputs fit the patient's dynamic range of electrically evoked hearing.

4. PERFORMANCE

MATLAB's Filter Design and Analysis tool (FDA Tool) is used to generate Band Pass FIR filter coefficients by using Hamming window of 128 orders. The magnitude response of the designed 8 band FIR filters as shown in Figure 7 with corresponding frequency bands in the Table 1.

Figure 6: Software Implementation of 8 Channel CIS Algorithm

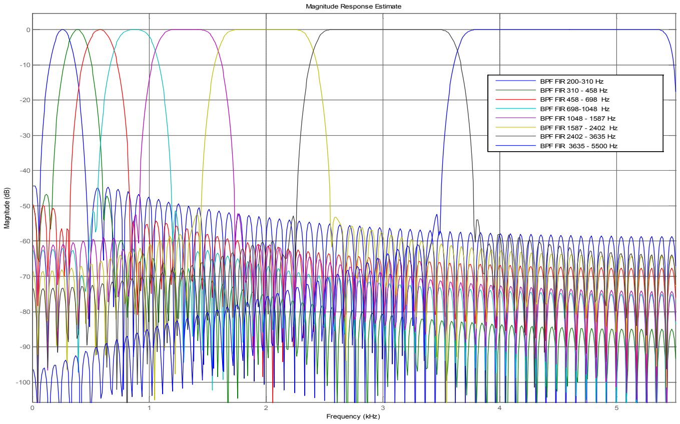


Figure 7: Magnitude Response of 8 Channel Band Pass Filter

Table 1: Filter Bands for 8-Channel CIS Algorithm.

Band Number	Frequency in Hz	Center Frequency In Hz
1	200-303	251
2	303 - 458	380
3	458 - 693	575
4	693 -1049	871
5	1049 -1587	1318
6	1587 -2402	2000
7	2402 -3635	3018
8	3635 -5500	4570

The fixed 8-channel CIS algorithm is modified in our work to suit to either 4-channel CIS or 5-channle CIS or 6-channel CIS or 7-channel or 8-channel operation with flexibility in programming n-channel CIS algorithm where n = 4, 5,6,7,8 that is highly needed based on selected number (4/5/6/7/8) of active electrode out of 12 electrodes of electrode array placed in cochlea [Frijns, 2003]. The CIS algorithm is validated by using various single tone frequency signals generated by using signal generator. The input and output signals are observed at various stages of implemented CIS algorithm by using FPGA in 8-channel Scope Coder measuring instrument. The designed system can be programmed to perform 4 to 8 channel CIS algorithm based on the number of active electrodes for speech/sound frequency range 200 to 5500Hz. The ADSP2185 can perform (i) 8 channel, (ii) 7 channel, (iii) 6 channel, (iv) 5 channel and (v) 4 channel CIS algorithm. The frequency distribution for 4 to 8 channel for the allowable frequency range is shown in Table 2.

A total of (8+7+6+5+5=31) FIR band pass filters are designed to perform the required functionality. All 32 FIR bandpass filter performance is tested for functionality using single tone frequency signal from signal generator to the speakers. The testing of FIR bandpass filters is done as follows. The frequency of the signal generator is set to the each bands center frequency. For example, 8 channels BPF, the function generator frequency is set at 610Hz and given to the line input of CODEC and each channel BPF FIR filter output is sent to the Line output of CODEC and the output response are observed. The Observed waveforms for the frequency of 610 Hz frequency it belongs to band 3. As only band 3 FIR filter allows the signal, the remaining Band Pass Filters attenuates the signal. The recorded waveform for the 3rd Channel BPF FIR Filter of 8-channel CIS algorithm as shown in Figure 8. Full-wave Rectification: After each band pass filtering has done, the filter waveform is rectified by using full wave rectification. Sample output of 3rd channel output after rectification as shown in Figure 9. After rectification again the processed sample is filtered by the 32 order low pass filter of 0-400Hz to extract the temporal information of each channel. Output of the LPF FIR filter output is shown in Figure 10. After low-pass filtering the processed sample is stored in the buffer.

Table2: Frequency Distribution of 200 – 5500 Hz Frequency for 4/5/6/7/8 Channel CIS Algorithm

	Channel numbers							
Number of active Channels	1st Band (Hz)	2nd Band (Hz)	3rd Band (Hz)	4th Band (Hz)	5th Band (Hz)	6th Band (Hz)	7th Band (Hz)	8th Band (Hz)
4	200 - 458	458 -1050	1050-2400	2400-5500				
5	200 - 388	388 – 753	753 – 1460	1460-2835	2835-5500			
6	200 - 347	347 - 604	604 – 1049	1049 – 822	1822 - 3166	3166 - 5500		
7	200-321	321-516	516 - 828	828 – 1329	1329 - 2134	2134 - 3426	3426 - 5500	
8	200-303	303 - 458	458 - 693	693 -1049	1049 - 1587	1587 - 2402	2402 - 3635	3635 - 5500

Figure 8: Input and Output Signals of Channel 3 FIR Band Pass Filter with Mono Tone Frequency of 610 Hz

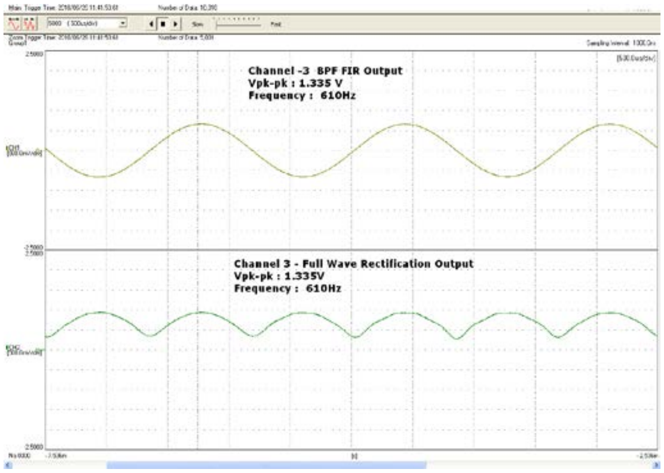

Channel -3 BPF FIR Output
Vpk-pk : 1.335 V
Frequency : 610Hz
Channel 3 - Full Wave Rectification Output
Vpk-pk : 1.335V
Frequency : 610Hz

Figure 9: Input and Output Signals of Channel 3 Full Wave Rectifier

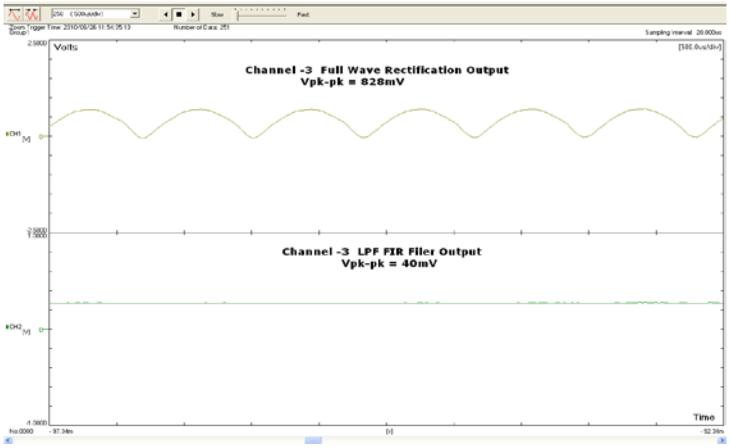

Volts
Channel -3 Full Wave Rectification Output
Vpk-pk = 828mV
Channel -3 LPF FIR Filer Output
Vpk-pk = 40mV
Time

Figure 1: Input and Output Signals of Channel 3 FIR LPF Filter (0-400Hz).

After rectification again the processed sample is filtered by the 32 order low pass filter of 0-400Hz to extract the temporal information of each channel. Output of the LPF FIR filter output is shown in Figure 10. After low-pass filtering the processed sample is stored in the buffer.

5 . CONCLUSION

The Configuration of FPGA for speech/sound processing has been particularly use with Bio-computing systems. The flexibility and computational power of the developed system allows speech processing using CIS strategy, which is tested and evaluated. Speech processing schemes may be improved by including surrounding noise pollutions, increase the no of channels and speech intelligibility optimization.

6. ACKNOWLEDGEMENTS

The funding and technical support provided by NSTL, Visakhapatnam is acknowledged.

7.REFERENCES

[1] [Clark, 2006] Graeme M. Clark, "The multiple-channel cochlear implant: the interface between sound and the central nervous system for hearing, speech, and language in deaf people—a personal perspective", Phil. Trans. R. Soc. B , 2006, Vol. 361,pp 791–810

[2] [[Zeng 2004] Fan-Gang Zeng,"Trends in Cochlear Implants", Trends In Amplification, Volume 8, No. 1, 2004, pp T1-T34.

[3] [Hirshorn, 1986] Michael S . Hirshorn, Dianne J. Mecklenburg, Judith A. Brimacombe, "Nucleus 22-channel cochlear implant: Preliminary observations", Journal of Rehabilitation Research and Development, April 1986, Vol . 23, No . 2, pp 27-33.

[4] [Hmida, 2007] Ghazi Ben Hmida, Hamadi Ghariani and Mounir Samet , "Design of Wireless Power and Data Transmission Circuits for Implantable Biomicrosystem", Biotechnology, Asian Network for Scientific Information , Vol 6, no. 2, 2007, pp 153-164.

[5] [Hochmair, 2006], Ingeborg Hochmair, Peter Nopp, Claude Jolly, Marcus Schmidt, Hansjörg Schößer, Carolyn Garnham and Ilona Anderson, "MED-EL Cochlear Implants: State of the Art and a Glimpse Into the Future", Trends in Amplification , Vol. 10, No. 4, December 2006, pp 201-220.

[6] [Wilson, 2007], Blake S. Wilson and Michael F. Dorman, "The Surprising Performance of Present-Day Cochlear Implants", IEEE Transactions on Biomedical Engineering, vol. 54, no. 6,pp . 969-973, June 2007

[7] H. J. McDermott, "A programmable sound processor for advanced hearing aid research," IEEE Trans Rehab. Eng., vol. 6, pp. 53-59, March 1998.

[8] Chih-Kuo Liang, Gin-Shu Young, Jia-Jin Jason Chen* and Chung-Kai Chen, "Microcontroller-based implantable Neuromuscular stimulation system with Wireless power and data transmission for Animal experiments, Journal of the Chinese Institute of Engineers, Vol. 26, No. 4, pp. 493-501,2003

[9] G. A. Kendir, W. Liu, G. Wang, M. Sivaprakasam, R. Bashirullah, "An optimal design methodology for inductive power link with class-E amplifier," IEEE Trans. Circuits Syst.I, Reg. Papers, vol. 52, no. 5, pp. 857–866, May 2005

[10] HUGH McDERMOTT., "An Advanced Multiple Channel Cochlear Implant ", IEEE Trans Bio Medical Engg, vol 36 pp 787-797,July 1989

[11] Shannon, R., Zeng, F-G., Kamath, V., Wygonski, J. and Ekelid, M., "Speech recognition with primarily temporal cues," Science, Vol.270 ,p303-304. October 1995.

[12] B. S.Wilson, C. C. Finley, D. T. Lawson, R. D.Wolford, and M. Zerbi, "Design and evaluation of a continuous interleaved sampling (CIS) processing strategy for multichannel cochlear implants," J Rehabil. Res. Dev., vol. 30, no. 1, pp. 110–116, 1993.

Comparative Study of 3G and 4G in Mobile Technology

K. Kumaravel

Assistant Professor
Dept. of Computer Science, Dr. N.G.P. Arts and Science College, Coimbatore, India – 641 048

Abstract—Mobile communication is one of the hottest areas and it is developing extremely fast in present times, thanks to the advances of technology in all the fields of mobile and wireless communications. Nowadays the use of 3G mobile communication systems seem to be the standard, while 4G stands for the next generation of wireless and mobile communications. This comparative study between 3G & 4G tells about the background and the vision for the 4G. We first present a review on the development history, characteristics, status of mobile communication and related 3G - 4G perspectives. An overall 4G framework features, having the basic keys (diversity and adaptability) of the three targets (terminals, networks, and applications). We present it in both external and internal diversity of each target to illustrate the causes and solutions of the adaptability feature. Then, the 4G domain of each feature in the framework is discussed from technical point, showing techniques and possible research issues for sufficient support of adaptability. At the end, a summary on 4G visions and some of the issues this new technology may face.

Keywords: OFDM, HSPA, LTE, MIMO, MC-CDMA, WCDMA, UMB

I.INTRODUCTION

Mobile broadband is becoming a reality, as the Internet generation grows accustomed to having broadband access wherever they go, and not just at home or in the office. Out of the estimated 1.8 billion people who will have broadband by 2012, some two-thirds will be mobile broadband consumers — and the majority of these will be served by HSPA (High Speed Packet Access) and LTE (Long Term Evolution) networks. People can already browse the Internet or send e-mails using HSPA-enabled notebooks, replace their fixed DSL modems with HSPA modems or USB dongles, and send and receive video or music using 3G phones. With LTE, the user's experience will be even better. It will further enhance more demanding applications like interactive TV, mobile video blogging and advanced games or professional services.

LTE offers several important benefits for consumers and operators: Performance and capacity - One of the requirements on LTE is to provide downlink peak rates of at least 100Mbit/s. The technology allows for speeds over 200Mbit/s and Ericsson has already demonstrated LTE peak rates of about 150Mbit/s. Furthermore, RAN (Radio Access Network) round-trip times shall be less than 10ms. In effect, this means that LTE — more than any other technology — already meets key 4G requirements.

II. DIFFERENTIATION BETWEEN 3G & 4G

3G is currently the world's best connection method when it comes to mobile phones, and especially for mobile Internet. 3G stands for 3rd generation as it just that in terms of the evolutionary path of the mobile phone industry. 4G means 4th generation. This is a set of standard that is being developed as a future successor of 3G in the very near future.

The biggest difference between the two is in the existence of compliant technologies. There are a bunch of technologies that fall under 3G, including WCDMA, EV-DO, and HSPA among others. Although a lot of mobile phone companies are quick to dub their technologies as 4G, such as LTE, WiMax, and UMB, none of these are actually compliant to the specifications set forth by the 4G standard. These technologies are often referred to as Pre-4G or 3.9G.

4G speeds are meant to exceed that of 3G. Current 3G speeds are topped out at 14Mbps downlink and 5.8Mbps uplink. To be able to qualify as a 4G technology, speeds of up to 100Mbps must be reached for a moving user and 1Gbps for a stationary user. So far, these speeds are only reachable with wired LANs.

Another key change in 4G is the abandonment of circuit switching. 3G technologies use a hybrid of circuit switching and packet switching. Circuit switching is a very old technology that has been used in telephone systems for a very long time. The downside to this technology is that it ties up the resource for as long as the connection is kept up. Packet switching is a technology that is very prevalent in computer networks but has since appeared in mobile phones as well. With packet switching, resources are only used when there is information to be sent across. The efficiency of packet switching allows the mobile phone company to squeeze more conversations into the same bandwidth. 4G technologies would no longer utilize circuit switching even for voice calls and video calls. All information that is passed around would be packet switched to enhance efficiency.

1. 3G stands for 3rd generation while 4G stands for 4th generation.
2. 3G technologies are in widespread use while 4G compliant technologies are still in the horizon.
3. 4G speeds are much faster compared to 3G.
4. 3G is a mix of circuit and packet switching network while 4G is only a packet switching network.

III. Features of 3G

3G telecommunications, is a generation of standards for mobile phones and mobile telecommunication services fulfilling the International Mobile Telecommunications-2000 (IMT-2000) specified by the International Telecommunication Union.] Application services include wide-area wireless voice telephone, mobile Internet access, video calls and mobile TV, all in a mobile environment. To meet the IMT-2000 standards, a system is required to provide peak data rates of at least 200 kbit/s. Recent 3G releases, often denoted 3.5G and 3.75G, also provide mobile broadband access of several Mbit/s to smart phones and mobile modems in laptop computers.

The following standards are typically branded 3G:

- the UMTS system, first offered in 2001, standardized by 3GPP, used primarily in Europe, Japan, China (however with a different radio interface) and other regions predominated by GSM 2G system infrastructure. The cell phones are typically UMTS and GSM hybrids.

Several radio interfaces are offered, sharing the same infrastructure:

- The original and most widespread radio interface is called W-CDMA.
- The TD-SCDMAradio interface, was commercialised in 2009 and is only offered in China.
- The latest UMTS release, HSPA+, can provide peak data rates up to 56 Mbit/s in the downlink in theory (28 Mbit/s in existing services) and 22 Mbit/s in the uplink.

IV. Features of 4G

4G, a range of new services and models will be available. These services and models need to be further examined for their interface with the design of 4G systems. Figures 2 and 3 demonstrate the key elements and the seamless connectivity of the networks.

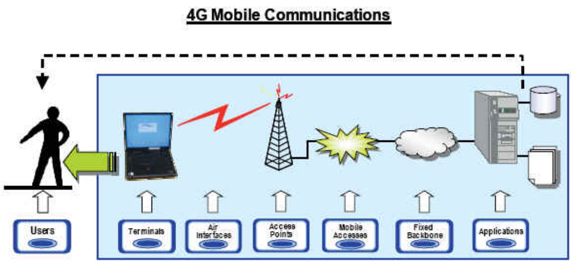


Figure 1. 4G Visions (Ref. 1)

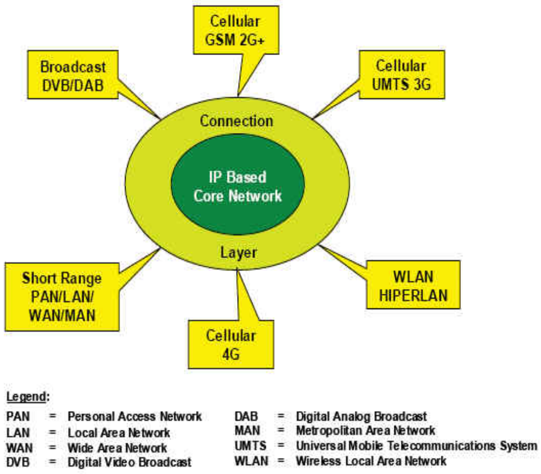

Cellular GSM 2G+
Broadcast DVB/DAB
Cellular UMTS 3G
Connection
IP Based Core Network
Layer
Short Range PAN/LAN/ WAN/MAN
WLAN HIPERLAN
Cellular 4G
Legend:
PAN = Personal Access Network
LAN = Local Area Network
WAN = Wide Area Network
DVB = Digital Video Broadcast
DAB = Digital Analog Broadcast
MAN = Metropolitan Area Network
UMTS = Universal Mobile Telecommunications System
WLAN = Wireless Local Area Network

Figure 2

4.1. Terminals

Till date the "terminal" for accessing mobile services has been the mobile phone. With the advanced 3G and also the 4G in future, we can also expect to see a broadening of this concept. User interfaces of terminals will vary from traditional keyboard, display, and tablet, to new interfaces based on speech, vision, touch, soft buttons, etc. These will be general-purpose computing and communication devices, and devices with more specific purposes to serve particular marker segments. There will still be recognizable mobile phones. But many of these will have larger screens to display Internet pages or the face of the person being spoken to. There will be smaller "smart-phones" with limited web browsing and e-mail capabilities. The addition of mobile communication capabilities to laptop and palmtop computers will speed up the Convergence of communication and computing, and bring to portable computing all the functions and features available on the most powerful desktop computers. There will be videophones, wrist communicators, palmtop computers, and radio modem cards for portable computers. Innovative new voice based interfaces will allow people to control their mobile communication services with voice commands.

4.2. Networks

Worldwide roll-out of 3G networks are delayed in some countries by the enormous Costs of additional spectrum licensing fees. In many parts of the world 3G networks do not use the same radio frequencies as 2G, requiring mobile operators to build entirely new networks and license entirely new frequencies. So that a number of spectrum allocation decisions, spectrum standardization decisions, spectrum availability decisions, technology innovations, component development, signal Processing and switching enhancements and inter-vendor cooperation have to take place before the vision of 4G will materialize.

4.3. Applications

The emerging applications for 3G and 4G wireless systems typically require highly Heterogeneous and time varying quality of service from the underlying protocol layers. So adaptability will be one of the basic requirements to the development and delivery of new mobile services. Promising techniques and possible topics may include: Mobile application should refer to a user's profile so that it can be delivered in a way most preferred by the subscriber, such as context-based personalized services. This also brings the applications with adaptability to terminals that are moving in varying locations and speeds. Techniques such as adaptive multimedia and unified messaging take the terminal characteristics into account and ensure that the service can be received and run on a terminal with the most suitable form to the host type.

The 4G technology will be able to support Interactive services like Video Conferencing (with more than 2 sites simultaneously), Wireless Internet, etc. The bandwidth would be much wider (100 MHz) and data would be transferred at much higher rates. The cost of the data transfer would be comparatively very less and global mobility would be possible. The networks will be all IP networks based on IPv6. The antennas will be much smarter and improved access technologies like OFDM and MC-CDMA (Multi Carrier CDMA) will be used. Also the security features will be much better.

Long-Term (Radio) Evolution or LTE is also part of 3G technology. It's a 3GPP its research item for Release 8. It's also known as 3.9G or "Super 3G" by some researchers. It's planned to commercialize in 2009. It was aims at peak data rates of 200 Mbps (DL) and 100 Mbps (UL).

The WiMax lobby and the people who are working with the WiMax technology are trying to push WiMax as the 4G wireless technology. At present there is no consensus among people to refer to this as the 4G wireless technology. I do not think this is popular with the researching community. WiMax can deliver up to 70 Mbps over a 50Km radius. As mentioned above, with 4G wireless technology people would like to achieve up to 1Gbps (indoors). WiMax does not satisfy the criteria completely. Also WiMax technology (802.16d) does not support mobility very well. To overcome the mobility problem, 802.16e or Mobile WiMax is being standardized. The important thing to remember here is that all the researches for 4G technology is based around OFDM. WiMax is also based on OFDM. This gives more credibility to the WiMax lobby who would like to term WiMax as a 4G technology. Since there is no consensus for the time being, we have to wait and see who would be the winner.

V. MULTIPLE ACCESS TECHNIQUES

3G wireless multiple access techniques are widely based on CDMA and WCDMA. But 4G demands a better multiple access technique for reducing the MAI (Multiple Access Interference) and ISI (Inter Symbol Interference) and thus improve the bit error rate performance. MC-CDMA is the best candidate that would satisfy the demands of 4G wireless systems. Moreover adaptive modulation techniques have been proposed for 4G, where the modulation scheme is changed dynamically based on the current channel estimates. MCCDMA is the hybrid combination of OFDM (Orthogonal Frequency Division Multiplexing) and CDMA. MC-CDMA with adaptive modulation promises to meet the demands of 4G regarding high data rate with a lower BER (Bit Error Rate).

OFDM has the capability to cancel multi-path distortion in a spectrally efficient manner. Rapid variation in channel characteristics are caused by multi-path and Doppler spread (due to the different speeds of mobile). Sometimes these time varying channels are characterized by very good SNR (Signal to Noise Ratio), but worse SNR at other times. So a fixed modulation technique cannot achieve the best Spectral efficiency as the system has to be built with a modulation scheme considering the worst case scenario. Hence during good channel conditions the system would not be able to obtain the best possible spectral efficiency. This is where adaptive modulation shows its role. Adaptive Modulation techniques takes advantage of the time varying channel characteristics and adjust the transmission power, data rate, coding and modulation scheme for the best spectral efficiency.

MC-CDMA

The basic idea of CDMA is to maintain a sense of orthogonality among the users in order to eliminate the MAI. This is done by employing orthogonal spreading codes to spread the data sequence. In MC-CDMA these spreading codes are defined in the frequency domain. Pseudo orthogonal codes can be used instead of orthogonal codes, thus increasing the number of users that can be accommodated. But pseudo orthogonal codes can increase MAI since the spreading codes are not fully orthogonal.

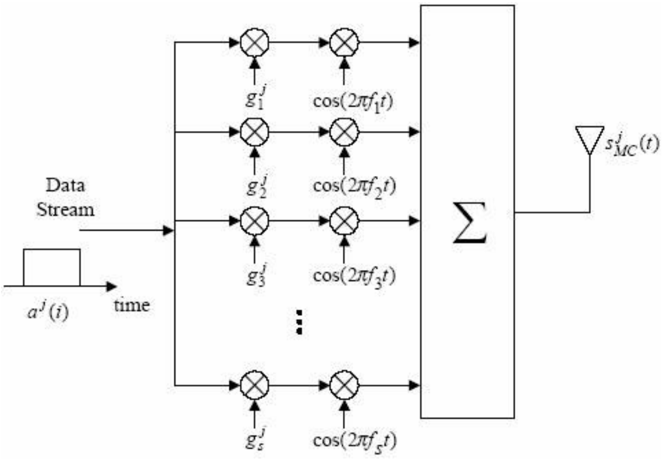


Figure 3: MC-CDMA Transmitter

Fig 3 shows the configuration of an MC-CDMA transmitter for user. It takes 3 the input data stream and converts into parallel data sequences each parallel data Sequence is multiplied with the spreading code. A guard interval in inserted between the symbols to eliminate ISI caused by multi-path fading.

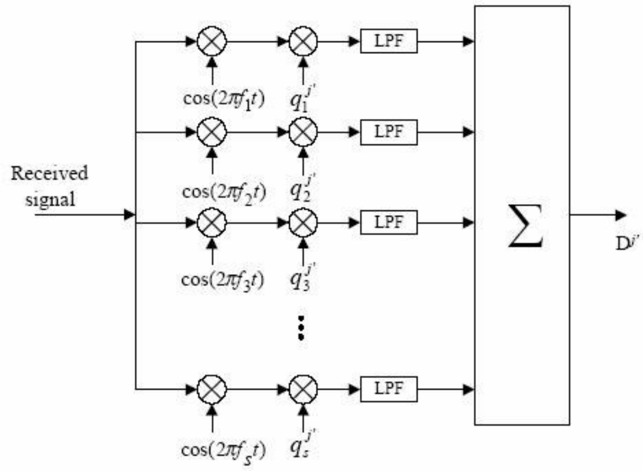


Figure 4: MC-CDMA receiver

In MC-CDMA receiver the received data are first coherently detected and then multiplied with the gain to combine the energy of the received signal scattered in the frequency domain. The system model for adaptive MC-CDMA is shown in the below fig 8.

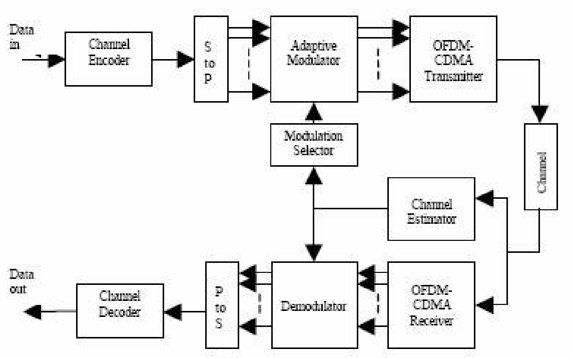


Figure 8: System model for adaptive MC-CDMA

VI. MULTIMEDIA – VIDEO SERVICES

4G wireless systems are expected to deliver efficient multimedia services at very high data rates. Basically there are two types of video services: bursting and streaming video Services. Streaming is performed when a user requires real time video services, in which the server delivers data continuously at a playback rate. Streaming has little memory requirement as compared to bursting. The drawback of streaming video is that it does not take advantage of available Bandwidth. Even if the entire system bandwidth is available for the user, streaming video service will transmit data only at a particular playback rate. Bursting is basically file downloading using a buffer and this is done at the highest data rate taking advantage of the whole available bandwidth. The flaw with this type of transmission is that it demands a large memory requirement. So work is being done to come up with a new scheme that limits the memory requirements and can exploit the available bandwidth of the system. The simulation details and comparison of streaming and bursting video comparison.

VII. Applications of 4G

Virtual Presence: This means that 4G provides user services at all times, even if the user is off-site. Virtual navigation: 4G provides users with virtual navigation through which a user can access a database of the streets, buildings etc of large cities. This requires high speed data transmission.

7.1.Tele-Medicine: 4G will support remote health monitoring of patients. A user need not go to the hospital instead a user can get videoconference assistance for a doctor at anytime and anywhere.
7.2. Tele-geoprocessing applications: This is a combination of GIS (Geographical Information System) and GPS (Global Positioning System) in which a user can get the location by querying.
7.3.Crisis management: Natural disasters can cause break down in communication

systems. In today's world it might take days or weeks to restore the system. But in 4G it is expected to restore such crisis issues in a few hours.

7.4. Education: For people who are interested in lifelong education, 4G provides a good opportunity. People anywhere in the world can continue their education through online in a cost effective manner.

VIII. Wi-Fi vs. WiMax

Comparing WiMax to Wi-Fi is akin to comparing apples to oranges. Initially it's easy to see why the comparison would exist, as most people think WiMax is merely a more robust version of Wi-Fi. Indeed they are both wireless broadband technologies, but they differ in the technical execution and ultimately their business case is very different. In addition to the technical differences that exist, the marketplace difference is that equipment is more or less non-existent for WiMax and certainly not geared towards a residential environment with very high pricing to be expected. It will take at least 2 years to see equipment of mass market uptake pricing.

WiMax could not be commercially available until the second half of 2005, and even then at a very controlled level. This is primarily due to standardization issues. In fact, it could not be until 2006 that a robust production and implementation would happen due to the ramp-up period for manufacturers. This is certainly one challenge to the widespread adoption of WiMax. Additionally, WiMax will have issues of pricing, and will remain far more expensive than Wi-Fi. WiMax will be primarily adopted by businesses to replace or displace DSL, and offices that want to cover a lot of territory without entering the world of endless repeaters that are necessary with the 802.11 technologies. It will take some time (2 years) for WiMax to significantly reduce its price-point for residential uptake. WiMax will not displace Wi-Fi in the home because Wi-Fi is advancing in terms of speed and technology. Each year brings a new variant to the 802.11 area with various improvements.

Additionally, for commercial deployment, frequency allocation will be an issue. With the three dominant communications players controlling the best frequencies, it will be hard to get the type of traction needed with the remaining companies operating in the frequencies available. WiMax will become extremely robust and displace Wi-Fi as the deployment of choice for commercial deployments, but that could not even begin until the end of 2006. Based upon the number of public hotspots already deployed, WiMax will not be chosen to replace those as they are up and running adequately and personnel involved understand how to work with the technology. The business case does not exist at the hotspot level. Where it may exist is for wider free use deployments such as city deployments (free ones) and other government sponsored or carrier sponsored (with ultra inexpensive pricing for consumers) deployments. If this happens then it's only Wi-Fi that will be displaced, but also cable and DSL will also lose a percentage of their subscriber base. What will cause the displacement is the consumer's proven desire for a bundled package.

IX. CONCLUSION

4G seems to be a very promising generation of wireless communication that will change the people's life to wireless world. There are many striking attractive features proposed for 4G which ensures a very high data rate, global roaming etc. New ideas are being introduced by researchers throughout the world, but new ideas introduce new challenges. There are several issues yet to be solved like incorporating the mobile world to the IP based core network, efficient billing system, smooth hand off mechanisms etc. 4G is expected to be launched by 2010 and the world is looking forward for the most intelligent technology that would connect the entire globe.

Someday 4G networks may replace all existing 2.5G and 3G networks, perhaps even before a full deployment of 3G. multiple 3G standards and springing up that would make it difficult for 3G devices to be truly global.

REFERENCES

1.B.G. Evans and K. Baughan, “Visions of 4G,” Electronics & Communication Engineering Journal, Vol. 12, No. 6, pp. 293–303, Dec. 2000.
2. C. R. Casal, F. Schoute, and R. Prasald, “A novel concept for fourth generation mobile multimedia communication,” in 50th Proc. IEEE Vehicular Technology Conference, Amsterdam, Netherlands, Sep. 1999, Vol. 1, pp. 381–385.
3. S. Y. Hui, K. H. Yeung, “ Challenges in the migration to 4G mobile systems,” Communications Magazine, IEEE , Volume: 41 , Issue: 12 , Dec. 2003, pp:54 – 59
4.A. Bria, F. Gessler, O. Queseth, R. Stridh, M. Unbehaun, J. Wu, J. Zander, “4th-generation wireless infrastructures: scenarios and research challenges,” Personal Communications, IEEE [see also IEEE Wireless Communications], Volume:8, Issue:6, Dec.2001, pp:25 – 31 [6] U. Varshney, R. Jain, “Issues in emerging 4G wireless networks,”Computer, Volume:34, Issue:6, June2001, pp:94 – 96
5. K. R. Santhi, V. K. Srivastava, G. SenthilKumaran, A. Butare, “Goals of true broad band's wireless next wave (4G-5G),” Vehicular Technology
Conference, 2003. VTC 2003-Fall. 2003 IEEE 58th , Volume: 4 , 6-9 Oct.
2003, Pages:2317 - 2321 Vol.4
6. L. Zhen, Z. Wenan, S. Junde, H. Chunping, “Consideration and research
issues for the future generation of mobile communication,” Electrical and Computer Engineering, 2002. IEEE CCECE 2002. Canadian Conference on , Volume:3, 12-15May,2002 , pp:1276 - 1281 vol.3
7. S. Chatterjee, W. A. C Fernando, M. K.. vasantha, “Adaptive modulation based MC-CDMA systems for 4G wireless consumer applications,” Consumer Electronics, IEEE Transactions on , Volume: 49 , Issue:4, Nov.2003, pp:995 – 1003
8. J. B. Chia, “Video services over 4G wireless networks: not necessarily Streaming,” Wireless Communications and Networking Conference, 2002. WCNC2002. 2002 IEEE , Volume: 1 , 17-21 March 2002 , pp:18 - 22 vol.1

Author:

Kumaravel Krishnan MCA., M.Phil.,CCNA., Serving as a Asst. Professor in Computer Science, Dr.N.G.P. Arts and Science College, Coimbatore. He has presented many papers in various conferences and published referred journals. He is pursuing Ph.D programme in computer science. He has more than a decade of experience in teaching and research.

Modelling An Enhanced Routing Protocol For Wireless Sensor Networks Using Implicit Clustering Technique

Idigo Victor[1], , Azubogu A.C.O[2], Oguejiofor Obinna3,Nnebe Scholar[4]

[1]Department of Electronics and Computer Engineering, Nnamdi Azikiwe University,Awka
Anambra State(234),Nigeria

[2]Department of Electronics and Computer Engineering, Nnamdi Azikiwe University,Awka
Anambra State(234),Nigeria

[3]Department of Electronics and Computer Engineering, Nnamdi Azikiwe University,Awka
Anambra State(234),Nigeria

Abstract: The localization of sensor nodes can be a very enabling technology that can help in improving the performance of many algorithms designed for wireless sensor networks (WSNs). This work is geared towards developing a positioning system that uses received signal strength based on fingerprinting technique. The proposed system models the signal strength distribution received from the sensor nodes using non-parametric or Gaussian distribution. The probabilistic Bayesian technique was employed as the localization algorithm for the basic model. The result obtained shows that there is an improved median error, appropriately 1.0 meter, compared to 2.5 meters for the nearest neighbour (NN) algorithm. Implicit clustering technique was used to enhanced the result obtain from the basic model. The performance of the basic model was enhanced by more than 18% for the static model and more than 10% for the mobile model. Finally using the enhanced model reduces the average number of operations per location estimate by more than 30%.

Keywords: Sensor nodes, Bayesian Technique, Clustering, Location estimate, Error probability

1.0. Introduction

Wireless sensor networks(WSN) are based on network of devices that can be densely deployed in an aggressive and inaccessible environments to sense the environment, and monitor with high accuracy the physical phenomena[1]. Each one of these devices is called a sensor node. They have limited processing speed, storage capacity and communication bandwidth.

In many WSN scenarios, the random deployment of hundreds of sensor nodes without localization hardware raises the problem of determining the topology of the network in terms of the outer boundary and the boundaries of the communication nodes [2]. Existing boundary recognition algorithms are able to determine these boundaries with certain accuracy. However, they only work for extremely dense networks and involve high computational and message complexities,

In multihop wireless networks, it is energy efficient to choose long paths along a series of short hops rather than short paths along a series of long hops. However, even though efficiency is always of paramount interest, it is not the only one [3]. Communications performance is also very important. By choosing many short hops we may lower the energy expenditure, but only to a certain degree; since delay increases, processing energy increases and control overhead increases. Therefore, the choice of how to incorporate energy is not as clear as it seems [3,4].

A useful distinction refers to whether energy is treated as a cost function or as a hard constraint [3], in the former case, the objective of the designer is to minimize the amount of energy per communication task, treating energy as an expensive but in exhaustive resource. However when energy is a hard constraint, the designer sees energy as a limited resource that will be exhausted[3,5]. In this case, the task is more complicated since there is a need to satisfy conflicting objectives; maximizing the longevity of the network versus communication performance such as throughput, total data delivery, etc.

The localization of sensor nodes can be a very enabling technology and can provide help to improve performance of many of the algorithms designed for WSNs. For example, in geographic routing protocols, the location information (in the form of coordinates) is used to select the next forwarding host among the senders neighbours. In rescue applications, rescue personnel can perform their task only if location of the hazardous event (reported by nodes) is known. Some related work in this area include the following: In [4 , 5] some methods for estimating unknown node positions using exclusively connectivity induced constraints are presented. These methods are only suitable for location determination with beacons. Some works reports are about an ingenious algorithm based on GPS free positioning. This algorithm explores only each node's knowledge of the neighbours and produces a coordinate system for each node and for the network. One major drawback of this approach is that the nodes do not know the physical direction of the coordinate system. In [6] the authors address the deployment by aircrafts of node groups and determine the positions of a node through the neighbours considering that the node is located close to the drop place of the node groups more represented. The bigger drawback is that the deployment does not always act like a model[4,6]. The authors use a mobile beacon to scan through the network, broadcasting its position while it passes. Although that is a good idea, it is not always possible to move beacons around a deployment area. [7] Addresses the problem of simultaneous localization, tracking and calibrations using probabilistic Bayesian filtering. This was reported to be a very good algorithm for ultrasound localization, but still lacks accuracy when using radio connections. However, this technique had been employed with some success in the field of robot localization [8]. In practice, the Bayesian localizer proves more accurate that the deterministic techniques such as the nearest neighbour(NN) algorithm because it takes into account more information from the training data during the data collection phase and filters the output using motion model.

This work employs the probabilistic Bayesian approach for the basic localization algorithm. The result obtained from the basic model was enhanced using the implicit clustering technique. Clustering of radio map locations was introduced as an approach to reduce the computational requirements of the location determination algorithm, improve accuracy and achieve scalability. The results show that using clustering reduces the average number of operations per location estimate by more than an order of magnitude.

2. Experimental Set up and Methodology

2.1 Experimental test bed

The test bed is located at the first floor of the 3-storey Administrative building of Nnamdi Azikiwe University, Awka. The floor has a dimension of 20m by 18m in an area of 360 sqm and segmented by a square of 1x1 meters as shown in fig 1. The deployment of the sensor nodes are shown marked AP1, ... AP4 in figure 1.

The transmitter and receiver were placed in different positions with respect to each other in the test bed. We used MSP430 mote which is developed by crossbow for the equipments. The mote employs the CC2420 ; which is a single-chip 2.4 GHz IEEE 802.15.4 RF transceiver with DSSS baseband modem of 2Mchips/s and 250kbps effective data rate with digital Received Signal Strength Indicator (RSSI), Link Quality Indicator (LQI) and MAC support. The transmitter nominal output was set to 0dBm and the receiver sensitivity was set at -90dBm. CC2420 has a built-in Received signal strength indication providing a digital value that can be read from the eight bit, signed 2's complement RSSI. RSSI_VAL register. The RSSI value is always averaged over eight symbol periods (128μs) in accordance with [9].The RSSI register value RSSI.RSSI_VAL can be referred to the power P at the RF pins by using the following equations;

P= RSSI_VAL + RSSI_OFFSET (1)

RSSI OFFSET is found empirically during system development and is approximately -45. For example, if reading a value of -20 from the RSSI register , the RF input power is approximately -65dBm.

The link quality Indication (LQI) measurement is a characterization of the strength and/or quality of a received packet as defined by [9]. Using the RSSI value directly to calculate the LQI value has the disadvantage that for example a narrowband interferer inside the channel bandwidth will increase the LQI value although it actually reduces the true link quality. CC2420 therefore also provides an average correlation value for each incoming packet. Software must convert the correlation value to the range 0-255 defined by [9], e.g. by calculating: LQI= (CORR-a).b . The Variables a and b are found empirically based on PER measurement as a function of correlation value.

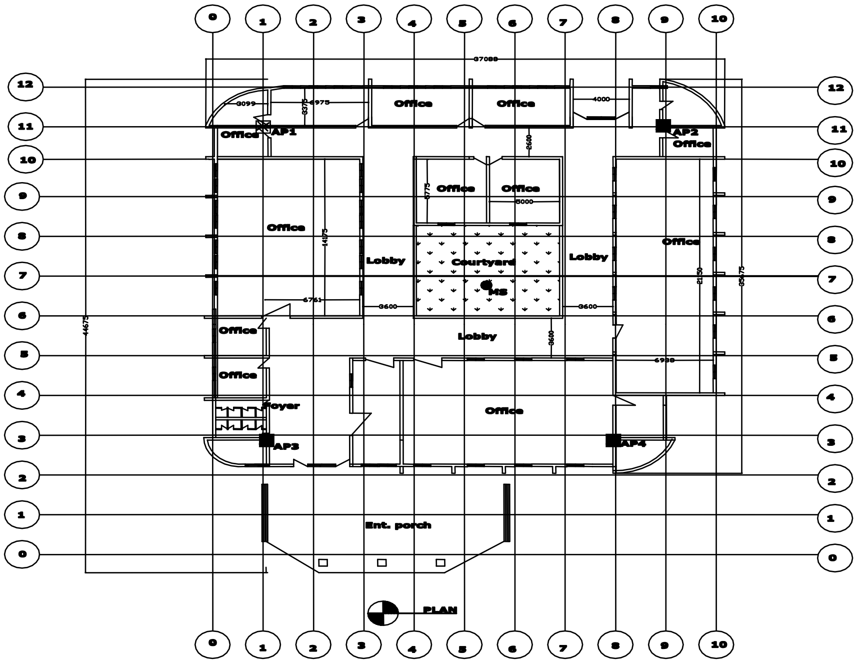


Figure 1: Experimental Test bed

2.0 Methodology

In the experiment , four different data packet sizes(20 bytes,30 bytes,50 bytes and 70 bytes) were transmitted over the wireless link of interest once every 100ms with transmitter-receiver separations of one meter. Ten measurements were taken for every T-R separation(fig 1). Ten such measurements were taken for each of the packet size but with 1-meter increment of the transmitter-receiver separation, up to 10 meters. In addition we sampled the Received signal strength indicator(RSSI) for every byte of data received. Averaging these values over an entire packet, an estimate of averaged received signal power for a packet was calculated.

This work is geared towards developing a positioning system that uses received signal strength based on finger printing method which is dependent on building a database. The reason for developing such a model is that today there is no way to develop and evaluate the performance of a positioning system except running massive measurements. Accuracy of the system is closely related to the number of nodes in the database and the distribution of them. In using fingerprinting method, it is required that a grid-network be built prior to any location estimation. After building the database for a new location , the new metric is measured irrespective of the viewed location and compared it with the database to find the best node, which could be referred to as desired point.

The proposed system models the signal strength distribution received from the sensor nodes using Gaussian distribution. The main advantage of the Gaussian technique is the efficiency of calculating the location estimate. [4,7,10] showed analytically that this technique is optimal among all discrete- space radio map-based location determination systems. The probabilistic Bayesian approach was employed as the localization algorithm for the basic model. The results obtained from the basic model is enhanced using the implicit clustering technique[11,12].

3. Problem Formulation and Implementation of Basic Model

Two vectors are normally used in estimating the location of the mobile station (MS). The first vector consists of samples of the RSSI measured at the mobile stations from N sensor nodes in the area. This Vector is denoted as $S = [s_1, s_2 , s_3 ..... s_n]$. The indoor positioning system estimates the mobile's location using the sample RSSI vector.

The second vector that forms the finger printing of the location, consists of the true means of all received signal strength random variable at a particular location from the N sensor nodes and recorded in the location database. We call it the location fingerprint or the average RSSI vector and denoted by $R = [r_1, r_2, r_{3,}.............r_n]$.

Let X be a 2 or 3 dimensional physical space. At each location xεX we can get the signal strength from N sensor nodes. The problem of the basic model becomes, given a signal strength vector $S(x) = [s_1, s_2 , s_3..... s_n]$. We want to find the location xεX that maximizes the probability $P({}^{x}/_{S})$.

T_o solve this problem , the probabilistic method of finger printing such as the Bayesian approach to WLAN localization was used.

This has been employed with some success in the field of robot localization. If I_t is the location at time t, 0_t is an observation made at t (the instantaneous signal strength values) and N, the normalization factor that ensures all probabilities sum to 1, then for localization , Bayes rule for static situation can be written as:

$$P(\frac{I_t}{0_t}) = P(\frac{0_t}{I_t}).\ P(I_t).\ N \qquad (2)$$

Equation (2) implies that the probability of being at location I given observation 0 is equal to the probability of observing 0 at location I , and being at location I in the first place. During localization, this conditional probability of being at location I is calculated for all finger prints. The most likely location is then the localizer's output. In order to calculate $P(\frac{I_t}{0_t})$ in equation (2), it is necessary to calculate the two probabilities on the right hand side of the equation. $P(\frac{0_t}{I_t})$ is known in Bayesian terms as the likelihood function. This can be calculated using the signal strength map. For each fingerprint, the frequency of each signal strength value is used to generate a probability distribution as a likelihood function.

Markov localization suggests using the transitional probability between locations. This probability is described as:

$$P(I_t) = P(\frac{I_t}{I_{t-1}})\, P(I_{t-1}) \qquad (3)$$

In other words, $P(I_t)$ is the sum of the transitional probability from all locations at t-1 to I at current time t , multiplied by the probability of being at these locations at t-1. $P(I_{t-1})$ is known from previous localization attempts. We calculate $P(\frac{I_t}{I_{t-1}})$ using a motion model, for instance, for a walking person the simplest and effective approach is to calculate the probability based on how far the user can move between t and t-1 [9]. The result obtained using Bayesian approach was compared to the use of nearest neighbour technique as reported in RADAR, an in-building RF based user location and tracking system[12].when 20 byte of data was transmitted figures (2) and (3) show the performance results using Bayesian and NN techniques for static and mobile localization, respectively. Static localization is performed for targets not expected to move and takes the prior probability as the uniform distribution. For mobile localization, the prior probability was calculated using a simple motion which caused the accuracy to be significantly improved compared to the NN approach . There is an improved media error when summarizing RSSI information as Gaussian approximately1.0 meter, compared to 2.5 meters for the NN in this test. In addition to improve accuracy the Bayesian localizer provide a frame work for the integration of other sensor nodes, infer red or mobile phone signal strength, can be integrated into the model by running the same Bayesian update equation on a shared probability distribution

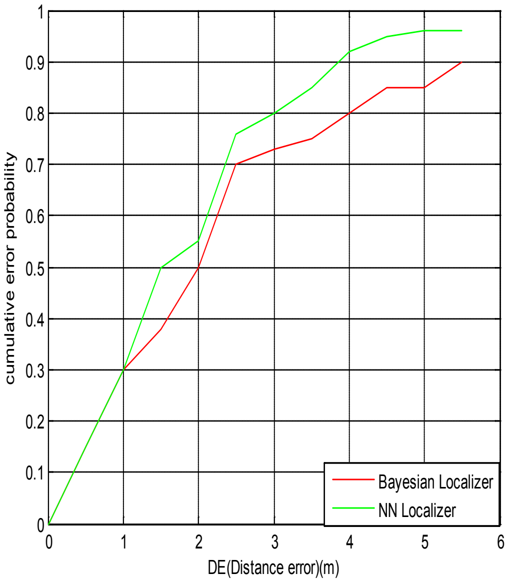


Fig 2: Cumulative Error probability for static localization

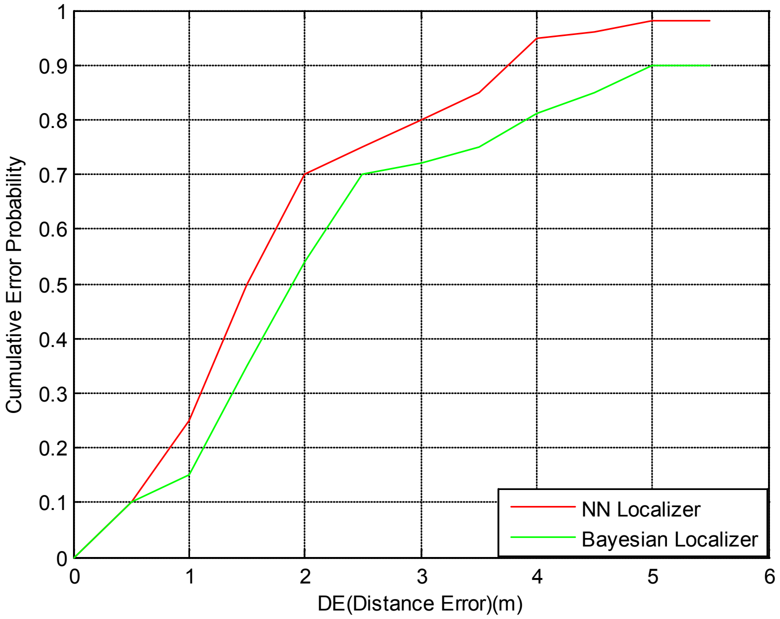


Fig 3: cumulative error probability for mobile localization

Figure 4 shows that the probability of error versus distance error for the four different data sizes. The results shows that the probability of error value is a function of the number of bytes transmitted.

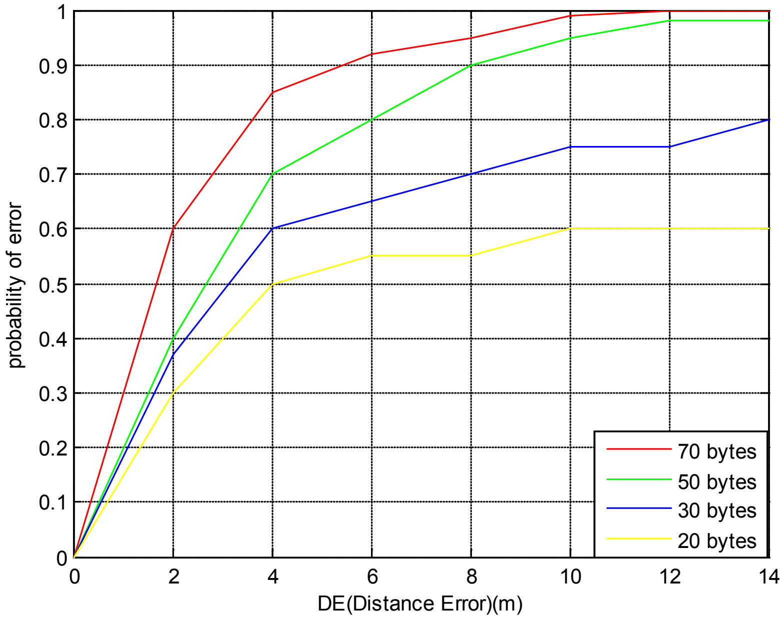


Figure 4: Comparative analysis of the probability of error versus Distance error for different data sizes

4. The Enhanced Model

Radio map location clustering was used in this work for the enhancement of the performance results defined by the basic model. This technique reduces the computational requirements of a WLAN location determination algorithm[5].

A cluster is defined as a set of locations sharing a common set of access points. This common set of access points is called the cluster key. The problem can be stated as : Given a location x, we want to determine the cluster if which x belong. Two clustering approaches are presented in [5]: explicit clustering where the system must determine the clusters during the offline or data collection phase as a separate step, and ,the implicit clustering where no special processing is performed in the offline phase but rather during the location determination phase, the system performs clustering implicitly. The explicit clustering technique was reported in [5,8] as producing slightly better accuracy than that of the implicit clustering technique. However, the average number of operations performed per location estimate for the clustering technique is much lower than the corresponding number of the explicit clustering technique [8].

In order to conserve energy during the location determination process, this work adopted the implicit clustering approach as the technique used in the enhanced model. The enhanced algorithm works as follows: considering a sequence of RSSI values from each sensor node , we start by sorting the sensor nodes in descending order according to their average RSSI values. Then for the node with the strongest average RSSI value , we calculate the probability of each location in the radio map set given observed RSSI sequence from this node alone. This gives us a set of candidate locations (location that have non-zero probability). If the probability of the most probable location is “significantly” higher (according to a threshold) than the probability of the second most probable location, we return the most probable location as the location estimate, after consulting only one node. If this is not the case, we go to the next node in the sorted sensor node list. For this node ,we repeat the same process again, but only for the set of candidate location obtained from the first sensor node. Finally, the algorithm returns the most probable location in the candidate list that remains after consulting all the nodes.

Figures (5) and (6) gives a comparative analysis of the cumulative error probability for the Nearest neighbour (NN), the basic model and the enhanced model algorithms. Figure (5) is the analysis for the static case while figure (6) for the mobile case.

Figure (7) shows the variation of the average number of operation per location estimate for the basic and enhanced models

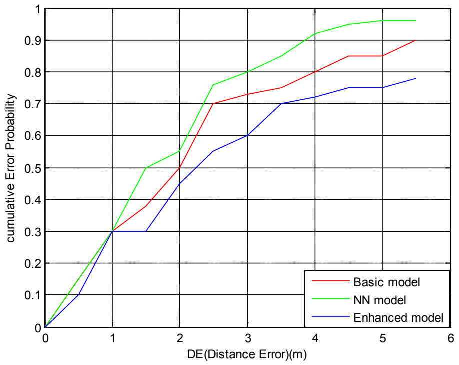


Figure 5: Cumulative Error Probability Analysis for NN, Basic and Enhanced models (Static Localization)

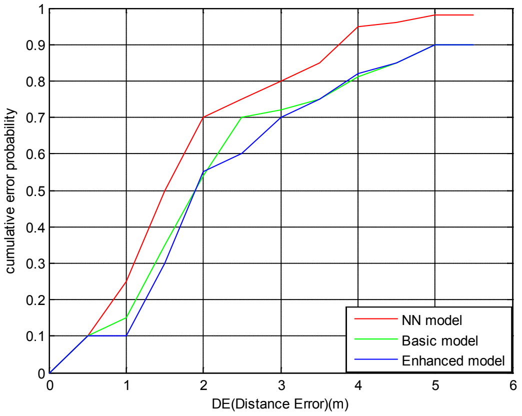


Figure 6: Cumulative Error Probability for NN, Basic, and Enhanced models(Mobile Localization)

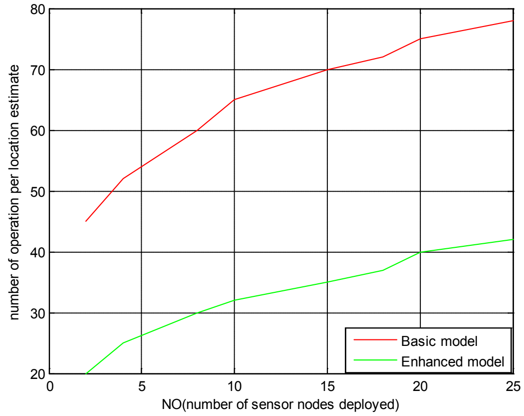


Figure 7: Comparative Analysis of number of operations required per location estimate versus the number of sensor nodes deployed

5. Conclusion

This work presents an empirical modelling of an enhanced indoor positioning system that uses RSSI/fingerprinting technique. The system models the signal strength distribution received from sensor nodes using non- parametric or Gaussian distribution. The main advantage of the Gaussian technique is the efficiency of calculating the location estimate.

The probabilistic Bayesian algorithm was employed for the basic model. The result obtained was compared to the NN algorithm used in RADAR. This is an improved median error when summarizing RSSI information Gaussian appropriately 1.0 meter, compared to 2.5 meters for the NN in this test. The performance analysis of the proposed model with respect to the data size transmitted was shown using probability of error technique, which is the probability that the location technique would give an incorrect estimate. Result shows that the performance of the system slightly decreases with increase in data size. The performance of the basic system was enhanced by more than 18% for the static model and more than 10% for the mobile model using the implicit clustering technique. Results also show that using the enhanced model reduces the average number of operations per location estimate by more than 30%.

References

[1]. Ming Zhang, Yanhang LU and Chenghong Gong, (2008), “Energy Efficient Routing protocol based on clustering and least square tree in Wireless sensor Networks”. International Conference on computer science and software Engineering pp 361 – 364.

[2]. Bager Zarei , Mohammed Zeynali and Vahid Majid Nezhad, (2010), IJCSI vol. 7, issue 4. www.ijcsi.org

[3]. Akyildiz I.F., weilians.,(2002), “A survey on Sensor Networks” IEEE communications Magazine, 40(8): 102-114.

[4]. Elgamal A.et al(2004), “A framework for monitoring Bridge and Civil

infrastructure" proceeding of 3rd China-Japan-US symposium structural health and control.

[5]. Cesare Alippi, Giuseppe Anastasi, Mario Di Francesco, Manuel Roveri (2009), " Energy Management in Wireless Sensor Networks with Energy-hungry Sensors". IEEE instrumentation and Measurement Magazine. Vol.12 no.2.

[6]. David Culler, Deborah Estrin, Mani Srivastavai(2004), "Overview of Sensor Networks", Special issue in sensor networks, IEEE computer 37(8), pp 41-49.

[7] Jan Rabacy, et al(2007), " Pico radio: communication/ computation piconodes for sensor networks". Technical Report, Electronic Research Laboratory, pp 7-22.

[8]. Kamik A. And Kumar A.(2004),"Iterative Localization in wireless adhoc sensor network : one dimension Case", in proceedings of the international conference on signal processing and communications.

[9]. IEEE std. 802.15.4 (2003): Wireless Medium Access Control(MAC) and Physical Layer(PHY) specifications for low rate wireless personal Area networks(LR-WPANs).

[10]. Monstafa Abdel A.V. (2004), "HORUS : A WLAN- Based Indoor location determining system", PhD dissertation, Worchester, polytechnic Institute.

[11]. Sichitiu M.L, Ramadurai V.(2003), "Localization of Wireless sensor networks with a mobile beacon: a mobile beacon based Bayesian approach to localizing network nodes.

[12]. Bahl P. And Padmanabhan V.(2000), "RADAR: An in-building RF-based user location and tracking system, proceedings of IEEE infocom, Tel-Aviv, Israel. Vol 2 pp 755-784.

Models of Growth Heterogeneous Cancer Cells with Chains Markoviens and Estimation of Their Fractal Dimension

Labib Sadek TERRISSA[1], Abdelhamid ZERROUG[2]

**[1] Department of Computer Science, University of Biskra,
Biskra, 07000, Algeria**

**[2] Departement of Mathematics Department, University of Biskra,
Biskra, 07000, Algeria**

Abstract

Although little work in biometrics uses fractal geometry, we will discuss here biometrics cancer tissue examined under a microscope or simulated. The main purpose of our work is the simulation of the heterogeneous growth of cancerous tumors and the analysis of the appearance of their textures. The problem is to quantify the irregularity of their edges, which help enormously oncologists to give diagnoses to evaluate the treatment issued to their patients.
We propose new algorithms, which generates growth models with the ability to produce a border irregularity similar to that of cancerous tumors and value their fractal dimension.
The established models have two types of parameters: Algorithms describing the structure, and Scalar to quantify aspects modeled

Keywords: Simulation, cancerous tumour growth, Markov fields, fractal dimension**.**

1. Introduction

Several methods have been proposed to simulate tumour growth as the approach used by cellular automata (Alarcon, T., H.M. Byrne, and P.K 2003) [2]. The models that we developed are based on the assumption that the tumour began as a single mother cell, which will gradually develop to form a cluster of girls cells .

Each daughter cell of the tumour could be linked to the mother by a connexity walk. Stochastic growth mechanisms have been carried out by Markov fields. To simulate a tumour epithelial monolayer, we used a grid plane. To characterize the form of compact cell clusters, we propose new algorithms, which generates growth models with the ability to produce a border irregularity Similar to that of cancerous tumours and value their fractal dimension [1].

The model we developed is done with help of a formal language to specify process of formation and evolution of structures random using carcinogenic among other Markov chains. The established models have two types of parameters: Algorithms describing the structure, and Scalar to quantify aspects modelled

Various attempts have been made to construct a mathematical model that describes tumour growth [6, 3], but the cases are too limited. Growth process dominated by surface diffusion and deposition were described in some deposits models (4-6)

2. Formulation of models

2.1 First model

The epithelial monolayer could be represented by a planar square grid A($m{\times}n$) whose elements Aij correspond to cells Cij. Each cell is connected by “adherent junctions” with 4 neighbours, and has a proper activity process which defines its different states, the eventual transformation from one type to another and the interaction with its neighbours. A given state of a cell at instant t+1 may change depending on its state and the states of the neighbours at instant t. A cell Cij may be ‘ill ’, in which case we set Aij = 1, or ‘healthy’, and Aij = 0.

Initially, all elements of A are set equal to 0 except one which is set equal to 1 at any position. This first element of ill cell IC corresponds to the mother cell engaged in a cancerous process.

From an ill cell IC, we generate a process which consists of visiting healthy cells in the four directions: left, right, up, down (Fig. 1). As shown in Fig. 2, we scan lines and columns in the order (1), (2), (3), (4) and stop scanning whenever an IC is encountered.

Three cases can occur: the actual visited healthy cell HC is surrounded by 1,2 or 3 ill cells in these directions. Hence, we introduce the following given probabilities α, β, γ: α (resp. β) (resp. γ) is the probability that HC falls ill when it is surrounded by one (resp. two) (resp. three) ill cell(s). HC cannot be surrounded by four ill cells.

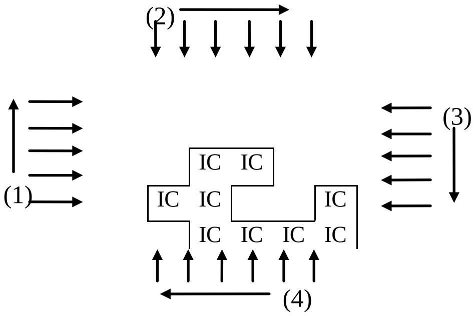


Fig. 1: visiting healthy cells.

The basic idea behind this model is that we do not visit the cell which has not orthogonal projection on the sides.

As in the initial state of the process there was only one cell sick, so there will be four sites to visit: Either A (i, j) = 1 the 1st cell disease.

So the four cells to visit are: A (i + 1, j), A (i – 1 , j), A (i , j + 1), A (i , j - 1)

NOTE:

whenever a rotation is established (i.e. go 1, 2, 3 and 4) see Fig1: we increase by two pixels each side, to visit the brink of a spot following the rotation

At each visit of a site that is naturally a healthy cell three scenarios are obtained. This implies the introduction of three probabilities:

Pr (C is sick / surrounded by 3 C patients) = γ

Pr (C is sick / surrounded by 2 C patients) = β

Pr (C is sick / surrounded by a sick C) = α

Then we draw a random variable Y = RND (1), and three cases this may present:

1st Case :

If γ [0, Y] then A (i, j) = 1 "is to say that the C is sick"

If not A (i, j) = 0.

2nd Case :

If β ε [0, Y] then A (i, j) = 1

If not A (i, j) = 0.

3 rd Case :

If α ε [0, Y] then A (i, j) = 1

If not A (i, j) = 0.

NOTE:

Aij is set at 1, but when the model is type (homogeneous) and Aij takes the values 1, 2, 3 or more when sick cell model is defined (Heterogeneous). According to the simulation model 1, α = 0.50 β = 0.55, γ = 0.75

Step 1 simulation based on probabilities

Step 2 Textures of the tumour

Step 3 Recovery of the border by the small number of squares (a minimum)

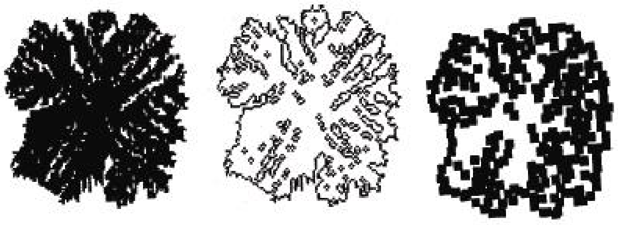


Fig. 2: cancer cells

2.2 Second model

The second model is to generate a process whose aim is to visit all the sites of the healthy state that are just boundary with the edge of the stain. With this model we got very few irregular spots , so far from approaching real cancerous tumor.

2.3 Third model

This model is identical to the 1st unless we introduce the following condition:

Each site can be visited only a single time, if it remains tests after the state healthy, it would no longer be visited another time. So we introduced an artifice to scoring, instead of leaving the site A (i, j) = 0 on the door at 2 in order to avoid the test a second time.

At the end of the simulation all sites at Level 2 will bring the state out.

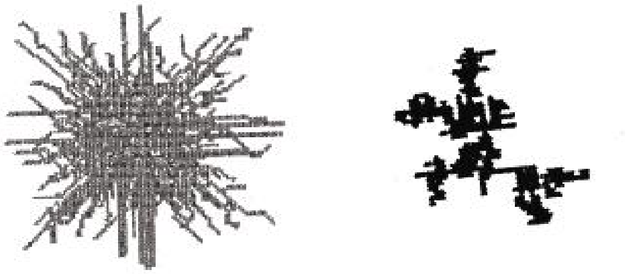


Fig. 3 : Simulation results (model 3)
α = 0.40, γ = 0.60, β= 0.80

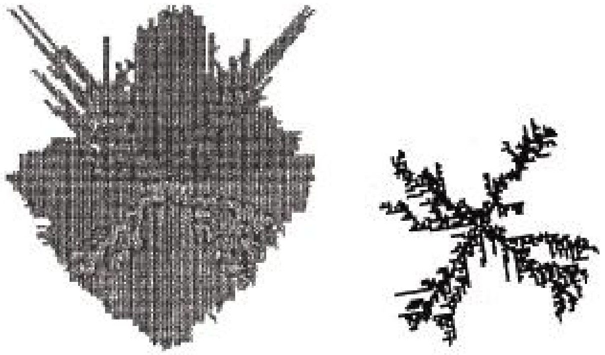


Fig. 4 : Simulation results (model 3)
α= 0.70, γ = 0.50, β = 0.60

3. Simulation of heterogeneous cancer tumors

In this paragraph it is assumed we have three types of cancer cells different C/c1; C/c2 and C/c3. This heterogeneity better reflects the reality of cancer in hospital environments. For this the process of the evolution of the tumour remains the same up to the stage or the test result is that the probabilistic test cell becomes sick. Then three possibilities may arise.

The test cell is surrounded by one, two or three sick cells. Next each case the cell test takes the nature of the cell number upper it, (Supi C / ci), i Є [1.3]. Which leads 19 scenarios to study for each test. The simulation steps of heterogeneous cancer tumors are presented by the organigram N°1.

4. Simulation with markoviens fields

Consider a region "S" shared flat (n * m) small squares called "pixels, which are located by couples (i, j) where i=1.. n and j=1..m.

4.1 Methodology

Let X be a field of Markov, with a value in a series of statements E, defined at all locations. In our case, E = { 0 , 1 } and S = { 1 , 2 ,..., n * m }

$$\text{Let} \qquad X = \begin{pmatrix} X_{11}\ldots\ldots\ldots\ldots X_{1n} \\ X_{21}\ldots\ldots\ldots\ldots X_{2n} \\ \\ X_{m1}\ldots\ldots\ldots\ldots X_{nm} \end{pmatrix}$$

J = Xij (state pixel A (i, j)). Hence Ω = { 0 , 1 }^{n*m} the total configurations The transition from one configuration X to another X * is performed on a field Markov dependence on local density P (x) which represents a priori distribution of X *.

4.2 Definition

A field is markovien local dependence if the state is taking the pixel A (i, j), depends only on the condition of neighboring pixels *A* (i, j). That is to say:

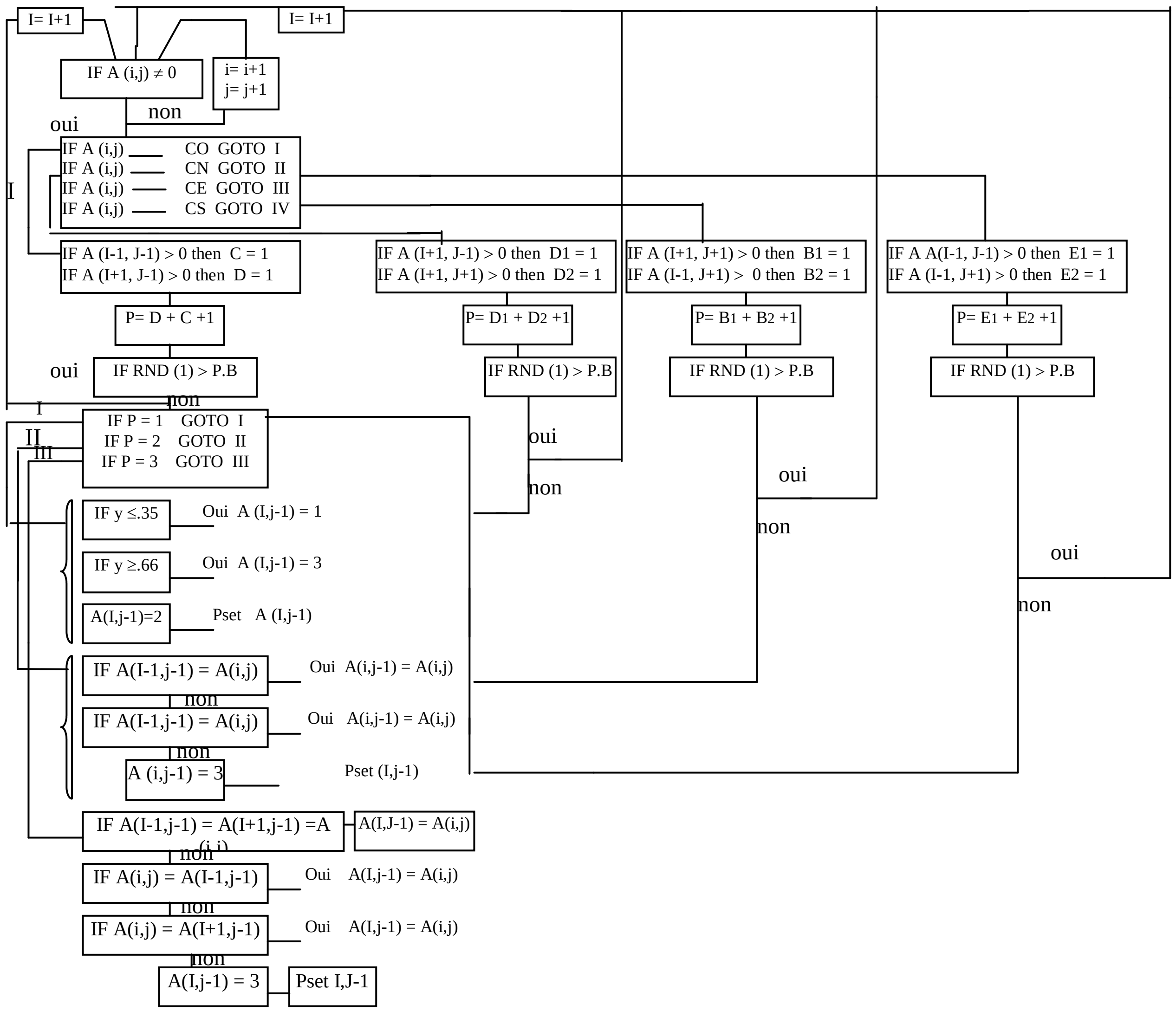
$$P\,(x\,(i,j)\,/\,x_m\,(i,j)\,) = P\,(x(i,j)\,/\,x_d\,(i,j))$$

Where x_m (i, j) represents the state of all the pixels other than A (i, j) and x_j (i, j) is all neighbours local A (i, j).

4.3 Markovien field of 1st order

Is submitted by: J (i, j) = closest neighbours from A (i, j).
If we consider two outcomes which differ only in the A pixel (i, j), we find that the conditional probability that the state appears K (i, j) (the rest being given), (ie holy and ill) only $\{0, 1\}$ Given that two states arise in our case, this means that we are facing a situation where states are disordered therefore a simple model is obtained by asking :

$$P\left(A_{ij} = {}^{k}/_{A_{\partial(i,j)}}\right) = \exp\left(B_k U_{ij}(k)\right) / \sum_k \exp\left(B_k U_{ij}(k)\right)$$



Organigram N°1 : The steps of simulation

ALGORITHM

1. Attribution of initial configuration of X.
2. Random visit all sites S.
3. Calculation on each site visited in the number of neighbours same state, and different state of the site.
4. Calculation probability of each state "K" in order to appear in (i, j) using (1)
5. Obtain a random variable Y and establish a test for each state "K":
   If Y ≤ P (x) the pixel takes the state K otherwise it is the second state to be selected.
6. Return.

The algorithm described above was done without the worry of the model that seeks to create (Task cancerous), but we realized that it does not accurately reflect the image of a task cancerous. That's why he has changed the mode of visiting the sites, which has prompted us to develop four modes of visits:

1st mode: The visit is made at random.
2nd mode: Visit column by column.
3rd mode: The visit is entering spiral.
4th mode: The visit takes place in orthogonal projection on the image while outgoing doing a spiral, we will explain later

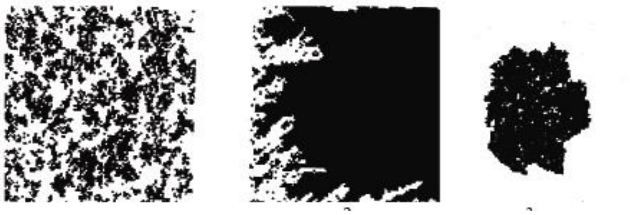


Fig. 5: Markovien model.

5. Evolution of the fractal dimension in the probabilities space

The size variation depends only on three parameters P1, P2, P3, it is estimated that it would be desirable to have an idea about the evolution of the dimension in the space formed by the probabilities P1, P2, P3. So we are given values to P1, P2, P3 so as to have a better spread over the whole space. Let P be in step with probability: p = 0.05 This gives us a distribution of space and generally fairly homogeneous, and that schematized as follows:

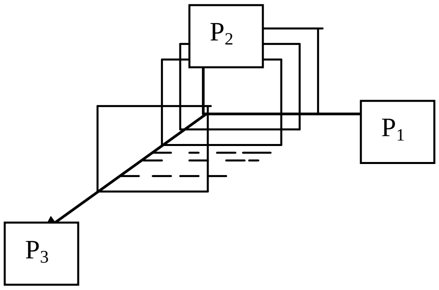


Fig. 6: Distribution of space.

We scans in the plane formed by P1, P2, P3, P1 induced space while respecting the assumption P1 <P2 <P3. Then for each triplet (P1, P2, P3) we do a 200 simulations (Randomization), which calculates the mean and standard deviation. as we are working on 80 points of space, which gives us 80 * 200 = 16000 simulations.

5.1 Data processing

The data acquired are two important points:

1. whenever P2, or P3 increases by a pitch (p = 0.05), the size decreases 0.02.
2. by cons p1 increases each time the same pitch p, the dimension increases 0.03.

The probability is constant, then one can conclude that the dimension is linearly dependent on the triple (P1, P2, P3) and is written:

$$D=a+ b.P1+c.P2+f.P3$$

Now it remains to find the four parameters a, b, c, f. so we make a multilinear regression. The 80 data we have: let $D(i) = a + b*P1(i) + c\,P2(i) + f\,P3(i)$

$$\begin{cases} a\sum_{i=1}^{80} P_1(i)+b\sum_{i=1}^{80} P_1^2(1)+c\sum_{i=1}^{80} P_1(i)P_1(i)+f\sum_{i=1}^{80} P_3(i)P_1(i)=\sum_{i=1}^{80} d_iP_1(i) \\ a\sum_{i=1}^{80} P_2(i)+b\sum_{i=1}^{80} P_2(i)P_1(i)+c\sum_{i=1}^{80} P_2^2(i)+f\sum_{i=1}^{80} P_3(i)P_2(i)=\sum_{i=1}^{80} d_iP_2(i) \\ a\sum_{i=1}^{80} P_3(i)+b\sum_{i=1}^{80} P_3(i)P_1(i)+c\sum_{i=1}^{80} P_2(i)P_3(i)+f\sum_{i=1}^{80} P_3^2(i)=\sum_{i=1}^{80} d_iP_3(i) \\ a\sum_{i=1}^{80} i+b\sum_{i=1}^{80} P_1(i)+c\sum_{i=1}^{80} P_2(i)+f\sum_{i=1}^{80} P_3(i)=\sum_{i=1}^{80} d_i \end{cases}$$

The Calculation of coefficient multilinear regression. with Statpal Regression software

5.2 Results (of 16000 simulations)

Dependent variable: Fractal Dimension.
Independent variable Model: P1, P2, P3.

Variable	Coefficients	Errors Std	Total score
Intersept	1.7488	0.0113	154.9098
P1	0.5060	0.0183	27.6554
P2	-0.6334	0.0170	-37.1914
P3	-0.4160	0.0163	-25.5748

So any tumors generates from P1, P2, P3 thier fractal dimension is estimated by:

$$D_F=1.7488+0,5060P1 – 0,6334P2 -0,4160P3$$

6. Conclusion

Once again the proposed method confirms its effectiveness for the following reasons:

- If P1 increases (even at (0.05)) the increases of the irregularity of the tumor cause an increase of the fractal dimension (detected by the method developed).
- if P2 and P3 increase, the irregularity decreases (even small) which causes the decrease in the fractal dimension, that is detectable by the method developed.

References

[1].A. Zerroug :S. Rebiai D. Schoëvaërt-Brossault : New Methods for Estimating the Dimension Fractal Introducing the

Artificial Intelligence ; Acta Appl Math .2008 DOI 10.1007/s10440-008-9358-4 Springer Science.

[2].Alarcon, T., H.M. Byrne, and P.K. Maini, A cellular automaton model for tumour growth in inhomogeneous environment. Journal of Theoretical Biology, 2003. 225(2): p. 257-274.
.

[3]. Gatenby, R.A ., Applications of competition theory to growth: implication for tumour biology and treatment Eur .J .Cancer 32A 722-726 1996.

[4]. Qi.A.S, Zheng, X, Du .C.Y., An, B.S.,:A cellular automaton model of cancerous growth. J.Theor.Biol 161.1-12 (1993).

[5]. Araujo, R.P. and D.L.S. McElwain, A history of the study of solid tumor growth: the contribution of mathematical modelling. Bulletin of Mathematical Biology, 2004. 66(5): p. 1039--1091.

[6]. Byrne, H.M. and M.A.J. Chaplain, Modelling the role of cell-cell adhesion inthe growth and development of carcinomas. Mathematical and Computer Modelling, 1996. 24: p. 1—17
.

[7]. Chen, W.Y., P.R. Annamreddy, and L.T. Fan, Modeling growth of a heterogeneous tumor. Journal of Theoretical Biology, 2003. 221: p. 205--227. 231--255.

[8]. Byrne, H.M., The role of mathematics in solid tumour growth. Mathematics Today, 1999.

Labib Sadek TERRISSA Holds an engineering diploma from Biskra University in electronic and Phd degree from university of le Havre, France. The Phd is about artificial retina modelisation and spike neuron study. We work no in the field of bioinformatic and neurocomputing. Currently he is the Vice dean of the faculty of science and life and associate prof in computer scence in Biskra university.

Abdelhamid ZERROUG Holds a Phd diploma from Biskra university (Algeria). His dissertation focused on the medical image processing and estimation the fractal dimension in dynamical systems. He is a an associate prof in mathematics department in Biskra University

Object Recognition Using Support Vector Machine Augmented by RST Invariants

R.Muralidharan[1], and Dr.C.Chandrasekar[2]

[1] **Assistant Professor,Department of Computer Applications, KSR College of Engineering, KSR Kalvi Nagar, Tiruchengode, Tamil Nadu 637215, India**

[2] **Associate Professor, Department of Computer Science, Periyar University, Salem, Tamil Nadu, India.**

Abstract

In this paper the support vector machine is utilized to recognize the object from the given image. The proposed method for object recognition is associated with the reduction of feature vector by Kernel Principal Component Analysis (KPCA) and recognition using the Support Vector Machine (SVM) classifier. Also in this paper the feature extraction method extracts features from global descriptors of the image. In the feature extraction process for an image, global features are extracted and formed as feature vector. For the entire training image the feature vector is generated and dimension reduction is done using KPCA. The reduced feature vector is used to train the SVM classifier. Later test images are given as input and tested the performance of the Classifier. To prove the efficiency of the SVM Classifier, Back Propagation Neural Network is used for the object recognition. From the comparison, SVM classifier outperforms.

Keywords: *Support Vector Machine, Object Recognition, Moment Inavariants, Kernel Principal Component Analysis.*

1. Introduction

Object recognition is a fundamental vision problem of identifying what is in the image, the concepts of object recognition has been applied in various fields like Manufacturing (for detecting defects/cracks in finished goods), surveillance system, optical character recognition, face recognition etc.,. The major task in object recognition is to identify if any, of an object from the set of known objects appear in the given image or image sequence. It plays an important role in Pattern recognition/classification and its key issues are whether selected features are stable and have good ability to differentiate different kinds of Objects. Object Recognition has been the focus of considerable research during the last four decades.

Hu was the first to introduce the geometric Moment Invariants which are invariant under change of size, translation, and orientation [Hu (1962)]. Since then many of the researchers had proposed moment invariants as pattern sensitive features in classification and recognition applications. Moments and functions of moments can provide characteristics of an object that uniquely represents its shape and have extensively employed as the invariant global features of an image in pattern recognition and image classification since 1960's.

[Borji (2007)] utilizes Support Vector Machine for recognition of Persian Font Recognition. [Chun-Jung Hsu (2001)] suggests Moment Invariants as feature for airport pavement distress image classification. [Rajasekaran (2000)] augmented the use of moment invariant as feature extractor for ARTMAP image classification. [Krishna (2010)] uses the support vector machine with the local features for classifying the leaf images. [Xin-Han (2010)] suggests that the support vector machine performs well in identifying micro parts. [Daniel (2011)] uses the moment invariants and Gray level co-variance matrix for the war scene classification. [Ronald (2006)] in his paper uses the support vector machine for automatic identification of impairments on eye diagram.

This paper discusses a formulation of an object recognition model in recognizing an object using Support Vector Machine. The features for the recognition are extracted from Geometric Moment Invariants and some of the image properties. During training phase the features are generated and feature vector is constructed. The constructed feature vector falls into high dimensional data, further processing of high dimensional data is a time consuming process. For reducing the dimension of the data, dimensionality reduction process KPCA [Narayanan (2009)] is applied. Through dimension reduction the support vector is created, which is provided as input to the SVM classifier to test the test image in recognizing the object.

The rest of the paper is organized as follows. Section 2 gives an overview of Moment Invariants and Kernel Principal Component Analysis. A summary of Classifier like Support Vector Machine, and Back Propagation Neural Network are given in Section 3. Section 4 gives an outline of the proposed system. Section 5 illustrates the Experimental results and section 6 for the conclusion.

2. Moment Invariants and Kernel Principal Component Analysis

2.1 Moment Invariants

Moments and moment invariants play a very important role as features in invariant pattern recognition. The approach using invariant features appears to be the most promising and has been used extensively since 1970. Its basic idea is to describe the objects by a set of measurable quantities called invariants that are insensitive to particular deformations and that provide enough discrimination power to distinguish objects belonging to different classes.

Global invariants like moment invariants are much more robust than local invariants with respect to noise, inaccurate boundary detection and other similar factors when compared to other moment Invariants. Moment invariants were first introduced to the pattern recognition and image processing community in 1962 [4], when Hu employed the results of the theory of algebraic invariants and derived his seven famous invariants to the rotation of 2D objects.

The two-dimensional geometric moment (m) of order (p+q)th of a function f(x,y) is defined as

$$m_{pq} = \int_{a1}^{a2} \int_{b1}^{b2} x^{p} y^{q} f(x, y)\, dx\, dy. \qquad (1)$$

where p,q = 0,1,2,……∞ and x,y gives the location of the pixel in the image along x-axis and y-axis respectively and f(x,y) gives the intensity value at a particular location. Note that the monomial product $x^p y^q$ is the basis function for this moment definition. A set of n moments consists of all m_{pq}'s for p + q ≤ n, i.e., the set contains ½(n+1)(n+2) elements.

Using non-linear combinations of geometric moments, Hu derived a set of invariant moments, which has the desirable properties of being invariant under image translation, scaling and rotation. However the reconstruction of the image from these moments is deemed to be quite difficult.

The Moment invariants are very useful way for extracting features from two-dimensional images. Moment invariants are properties of connected regions in binary images that are invariant to translation, rotation and scale.

The normalized central moments (2), denoted by ηpq are defined as

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\gamma}}\ . \qquad (2)$$

$$\textit{where} \qquad \gamma = \frac{p+q}{2} + 1,\ \ p+q = 2,3,4.....$$

A set of seven invariants can be derived from the second and third normalized central moments. This set of seven moment invariants (3) to (9) is invariant to translation, rotation, and scale change.

$$\phi_1 = \eta_{20} + \eta_{02} \qquad (3)$$

$$\phi_2 = \left(\eta_{20} - \eta_{02}\right)^2 + 4\eta_{11}^2 \qquad (4)$$

$$\phi_3 = \left(\eta_{30} - 3\eta_{12}\right)^2 + \left(3\eta_{21} - \eta_{03}\right)^2 \qquad (5)$$

$$\phi_4 = \left(\eta_{30} + \eta_{12}\right)^2 + \left(\eta_{21} + \eta_{03}\right)^2 \qquad (6)$$

$$\begin{aligned}\phi_5 = {}& \left(\eta_{30} - 3\eta_{12}\right)\left(\eta_{30} + \eta_{12}\right)\left[\left(\eta_{30} + \eta_{12}\right)^2 - 3\left(\eta_{21} + \eta_{03}\right)^2\right] \\ & + \left(3\eta_{21} - \eta_{03}\right)\left(\eta_{21} + \eta_{03}\right)\left[3\left(\eta_{30} + \eta_{12}\right)^2 - \left(\eta_{21} + \eta_{03}\right)^2\right]\end{aligned} \qquad (7)$$

$$\begin{aligned}\phi_6 = {}& \left(\eta_{20} - \eta_{02}\right)\left[\left(\eta_{30} + \eta_{12}\right)^2 - \left(\eta_{21} + \eta_{03}\right)^2\right] \\ & + 4\eta_{11}\left(\eta_{30} + \eta_{12}\right)\left(\eta_{21} + \eta_{03}\right)\end{aligned} \qquad (8)$$

$$\phi_7 = \left(3\eta_{21} - \eta_{03}\right)\left(\eta_{30} + \eta_{12}\right)\left[\left(\eta_{30} + \eta_{12}\right)^2 - 3\left(\eta_{21} + \eta_{03}\right)^2\right]$$
$$+ \left(3\eta_{21} - \eta_{30}\right)\left(\eta_{21} + \eta_{03}\right)\left[3\left(\eta_{30} + \eta_{12}\right)^2 - \left(\eta_{21} + \eta_{03}\right)^2\right] \qquad (9)$$

2.1 Kernel Principal Component Analysis

Principal Component analysis is a classic linear technique in statistical analysis. Given a set of values, PCA finds eigenvalue/vector, using only second-order statistics, a smaller set where the feature are uncorrelated to each others. The nonlinear version of PCA, namely Kernel Principal Component Analysis (KPCA), is able to extract the high order statistics, thus provides more information from the original data set. Kernel principal component analysis is one of the fundamental tools for unsupervised nonlinear dimension reduction and feature extraction. It involves calculation of the eigenvalue decomposition or singular value decomposition of centered kernel data and is in search for orthogonal functions that optimize the kernel data scatter. Similar to linear PCA, it involves the following eigen decomposition

$$CKC = i\sum i^T \qquad (10)$$

Where, K is the kernel matrix with entries $K_{ij} = k(x_i,x_j)$, C is the centering matrix Eq.(11) given by

$$C = I - \frac{1}{N} HH^T, \qquad (11)$$

I is the NxN identity Matrix, $H=[111\ldots1]^T$ is an N x 1 vector, $I = [a_1, a_2, \ldots a_N]$ with $a_i = [a_{i1},\ldots a_{iN}]^T$ is the matrix containing the eigenvectors and $\sum = diag(\lambda_1,\ldots\lambda_N)$ contains the corresponding eigenvalues. To denote the mean of the Φ - mapped data by $\Phi = \frac{1}{N}\sum_{i=1}^{N}\Phi(X_i)$ and define the centered map Φ as:

$$\Phi(X) = \Phi(X) - \Phi \qquad (12)$$

From the above centered map Eq.(12), the kth orthonormal eigenvector of the covariance matrix is computed. Then projection of $\Phi(X)$ onto the subspace spanned by the first n eigenvectors is computed.

In this paper the kernel function used is the polynomial function as in Eq. (13):

$$k\left(x_{i,}x_j\right) = \left(x_i^T x_j\right)^p,$$

where p = 1 gives standard PCA. (13)

The following are the steps involved in computing KPCA in the original space:

1. Compute the Kernel Matrix : $K_{ij} = K(x_i,x_j)$.
2. Center K.
3. Diagonalize K_c and normalize eigenvectors:

$$\lambda_k\left(\alpha^k . \quad \alpha^k\right) \quad = \quad 1 \qquad (14)$$

4. Extract the k first principal components

$$\Phi(X)^k_{kpc} = \sum_{i=1}^{N}\alpha_i^k \left(\Phi(X_i) \quad . \quad \Phi(X)\right) \qquad (15)$$

3. Classifier

3.1 Support Vector Machine

Support Vector Machine is one of the supervised Machine Learning Technique, which was first heard during COLT-92 introduced by Vapnik, Boser, Guyon. Support Vector Machines are used for classification and regression; it belongs to generalized linear classifiers. SVM is a mostly used method in pattern recognition and object recognition. The objective of the support vector machine is to form a hyperplane as the decision surface in such a way that the margin of separation between positive and negative examples is maximized by utilizing optimization approach. Generally linear functions are used as a separating hyperplane in the feature space. For achieving better performance, several kernel functions are used such as polynomial function and radial-bias function, in this paper, polynomial function is used as kernel function. When using kernel functions, the scalar product can be implicitly computed in a kernel feature space.

For the proposed work, the system starts with training sample $\{(x_i, y_i)\}_{i=1}^{N}$, where the training vector is x_i and its class label is y_i. The proposed method aims to find the optimum weight vector w and the bias b of the separating hyperplane such that [Haykin (1999)]

$$y_i\left(w^T\varphi(x_i) + b\right) \geq 1 - \xi_i, \qquad \forall_i$$
$$\xi_i \geq 0, \qquad \forall_i \qquad (16)$$

with w and the slack variables ξ_i minimizing the cost function given below

$$\phi(w,\xi_i) = \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i \qquad (17)$$

Where the slack variables ξ_i represent the error measures of data, C is the value assigned to the errors, and $\varphi(.)$ is a kernel mapping which maps the data into a higher dimensional feature space.

3.2 Back Propagation Network

A Back-Propagation neural network (BPN) consists of at least three layers of units: an input layer, at least one intermediate hidden layer, and an output layer as shown in Fig. 1. Typically, units are connected in a feed-forward fashion with input units fully connected to units in the hidden layer and hidden units fully connected to units in the output layer. When a Back-Propagation network is cycled, an input pattern is propagated forward to the output units through the intervening input-to-hidden and hidden-to-output weights. The output of a Back-Propagation network is considered as a classification decision.

The purpose of using Back-Propagation neural network in this study is to adopt the characteristics of memorizing and referencing properties that recognize the testing 2D image feature. The input of the network is the feature information extracted from the image. And the target is the designated index of the object. When training the BPN, the input pattern (x1,x2) is fed to the network, through the hidden layers to the output layer. The output pattern is compared with the target pattern to find the deviation. These extracted features are continuously fed into a BPN and the network will self-adjust until a set of weights (V11, V12, V21, V22, y11 and y21) with specified error value. Then these weights are stored and used for recognition later on.

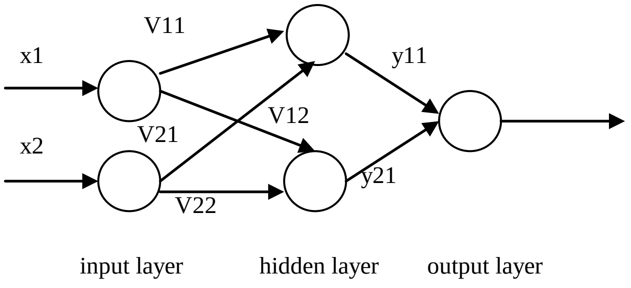


Fig. 1. Back-propagation neural networks

4. Proposed System

The proposed system for recognizing the object from an image is given below fig.2. The system is implemented as two phases. During first phase the system is trained with several set of training images. Initially the training images are preprocessed to obtain exact information for feature extraction. In the pre-processing stage, the noise in the image is removed and sharpening is done to reduce the effect of the illumination and lack of contrast on the training images. After pre-processing, the training image is applied for edge detection process, for edge detection Canny’s Edge detection method is used. In the feature extraction process, the moment invariants are extracted for each of the pre-processed training images. Feature extraction is defined as a process of converting the obtained image into a unique, distinctive and compact form. The computed moment invariants for all the pre-processed training images are arranged in such way to construct the feature vector. The constructed feature vector is high-dimensional. To reduce the high dimension of the data to low dimension without losing the important properties, Kernel Principal Component Analysis is done. This results into the support vector that can be used for the classifier Support Vector Machine. During testing phase, once the test image is given as input to the proposed system, the pre-processing and feature extraction process are done as specified in the training phase. The computed feature vector is given as input to the Support Vector Machine Classifier, based on the support vector generated during training phase; the input image is recognized and labeled. To compare the results of Support Vector Classifier, the Back Propagation Neural Network is trained with feature vector and tested.

5. Experimental Results

Table 1 and Table 2 show some of the selected results of our experiments. The experiment is conducted to estimate the recognition accuracy, and to verify the robustness of the proposed method. To experiment the proposed method, COIL-100 database which is widely used in 3D object Recognition researches [Nene (1996)]. This database consists of images of 100 different objects; each one is rotated with 5 degree angle interval in vertical axis. Hence for every object there are 72 images, which sum up to 7200 images for the whole database. The entire COIL-100 database is divided into two sets, one as training set and another one as test set. Three different training sets are formed for three different sampling angles (10, 30, and 50 degrees). The following fig.3 shows the set of sample images (Grayscale) used for forming the training set.

The proposed system is experimented with each set of training and test images and the results are shown in the table. The proposed method has been implemented using MATLAB. To evaluate the recognition accuracy for the proposed methods and the traditional methods the correct recognition percentages (CRP) were determined.

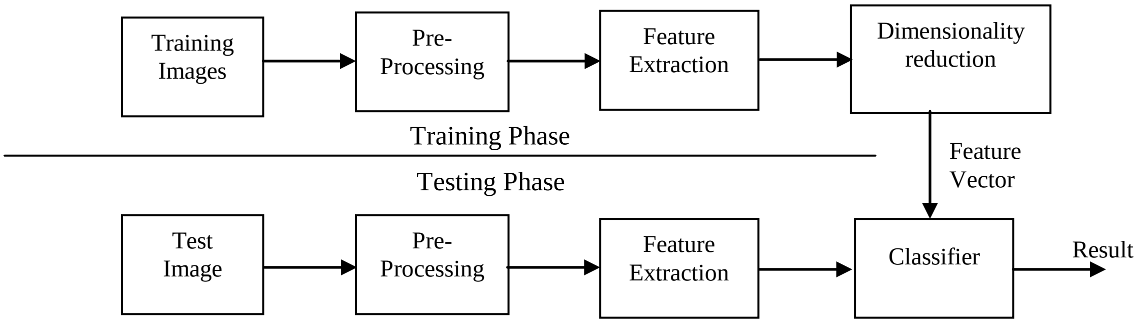


Fig.2. Schematic model of the proposed system

To verify the robustness, the proposed method is compared with performance of SVM using the original feature vector. The resultant values are shown in the table 1. and fig.4. The experiment is also conducted by increasing the number of training images. The resultant values for the increased training images are shown in the table 2. and fig.5. From the experimental results it is proved that the proposed method provides better performance in terms of CRPs compared with the traditional method. The CRP is computed as

$$CRP = \frac{Np}{T} \tag{18}$$

Where Np is number of positive recognition and T is total number of test conducted. The CRP values are provided in the table 1 and 2 and the performance is shown in the graph fig.4 and fig.5.

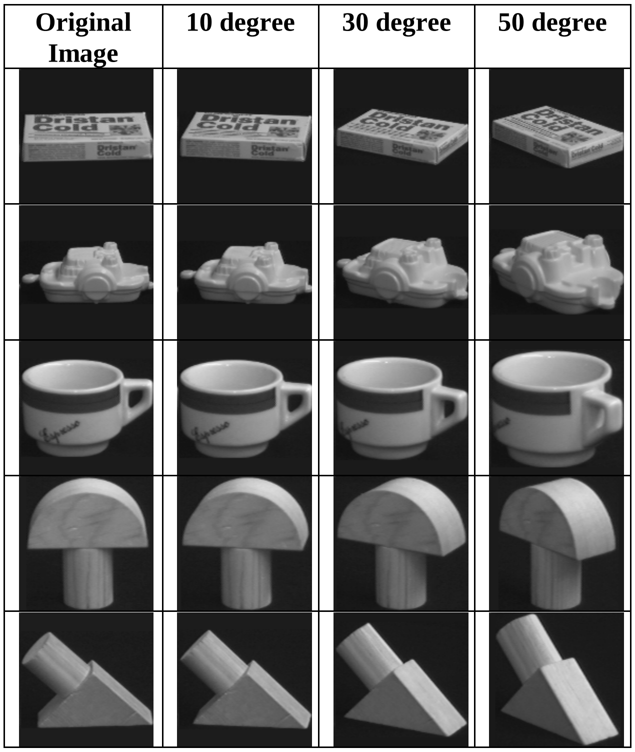


Fig.3. Sample images from COIL-100 Database

Table 1: Recognition performance in terms of CRPs of the proposed method, SVM and BPN.

Method	CRP %			
	10 degree	30 Degree	50 Degree	ALL (10, 30, 50)
SVM+KPCA (Proposed)	93.2	94	94.3	96.3
SVM	86.4	85.5	87.4	88.3
BPN	78	77.5	78.1	78.5

Table 2: Correct Recognition Percentage for different number of samples using proposed method, SVM and BPN.

Number of Samples	CRP%		
	SVM+KPCA	SVM	BPN
25	85.4	80	75
50	88.9	83.5	76.4
75	90.4	84.9	78.4
100	93.8	85.4	78.7
125	97.8	88.7	79.4

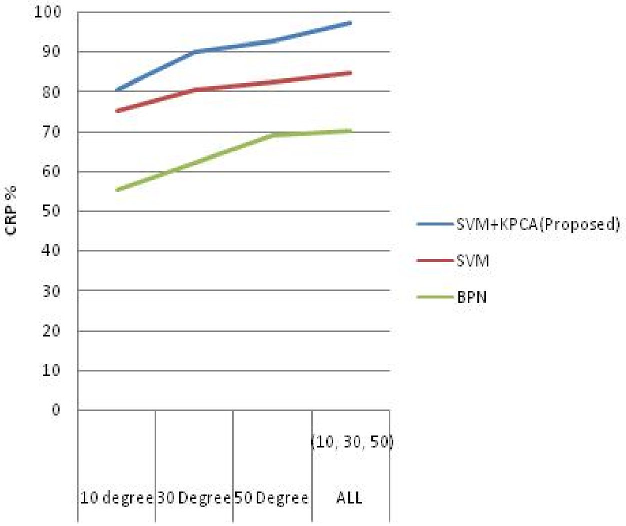


Fig.4 Recognition performance for the different training sets

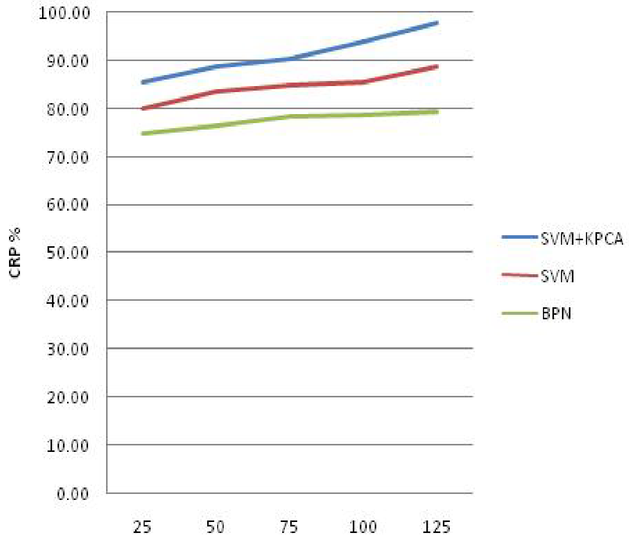


Fig.5. Correct recognition percentage for number of samples

6. Conclusion

This paper has presented SVM based object recognition using the Moment invariant Features. We have shown how the SVM recognizes the objects using the polynomial based kernel function. Also the KPCA is used for dimensionality reduction. For comparing the proposed method (SVM + KPCA) results, back-propagation method is implemented. Our proposed method is implemented in MATLAB with testing and training images available in the COIL-100 database. The SVM classifier performs well and provides high recognition rate compared to back-propagation network method.

References

[1] Borji, A.and Hamidi, M., “Support Vector Machine for Persian Font Recognition“, World Academy of Science, Engineering and Technology, Vol.28, 2007, pp.10 – 13.

[2] Chun-Jung Hsu, Chi-Farn Chen, Chau Lee, and Shu-Meng Huang, “Airport Pavement Distress Image Classification Using Moment Invariant Neural Network“, CRISP, 2001, pp. 123 – 127.

[3] Daniel Madan Raja, S., and Shanmugam, A. “Artificial Neural Network Based War Scene Classification using Invariant Moments and GLCM Features: A Comparative Study”, International Journal of Engineering Science and Technology, Vol.3,No.2, 2011, pp. 1189 – 1195.

[4] Duda, and Hart, Pattern Classification and Scene Analysis, Wiley, New York, 1973.

[5] Hu, M.K, “Visual Problem recognition by Moment Invariant”, IRE Trans. Inform. Theory, IT-8,1962, pp.179-187.

[6] Haykin S, Neural Network: A Comprehensive Foundation, Prentice Hall, New Jersey, 1999.

[7] Jan Flusser, Tomas Suk, and Barbara Zitova, Moments and Moment Invariants in Pattern Recognition, John Wiley & Sons, Ltd., 2009, ISBN:978-0-470-69987-4.

[8] Krishna Singh, Indra Gupta, and Sangeeta Gupa. “SVM-BDT PNN and Fourier Moment Technique for Classification of Leaf Shape”, International Journal of Signal Processing, Image Processing and Pattern Recognition, vol.3,no.4, 2010, pp. 67-78.

[9] Mi Hye Song, Jeon Lee, Sung Pil Cho, Kyoung Joung Lee, and Sun Kook Yoo, “Support Vector Machine Based Arrhythmia

Classification Using Reduced Features", International Journal of Control, Automation and Systems, vol-3, no-4, 2005, pp. 571 – 579.

[10] Muralidharan, R.; Chandrasekar, C. "Scale Invariant Feature Extraction for Indentifying an object in the image using Moment Invariants", Proceedings of IEEE International Conference on Communication and Computational Intelligence, 2010, pp.454 – 458.

[11] Narayanan Sundaram, Support Vector Machine Approximation using Kernel PCA, Technical Report, University of California, USA, 2009.

[12] Nene S.A, Nayar S.K, and Murase H, Columbia object image library (coil-100). Technical report, Colmubia University, 1996.

[13] Perantonis, S. J., and Lisboa,P.J.G, "Translation, rotation, and scale invariant pattern recognition by high-order neural networks and moment classifiers", IEEE Trans. Neural Networks, Vol-3, 1992, pp. 241–251.

[14] Rajasekaran, S., and Vijayalakshmi Pai, G.A. "Image recognition using Simplified Fuzzy ARTMAP augmented with a moment based feature extractor", International Journal of Pattern Recognition and Artifical Intelligence, Vol-14, No-8, 2000, pp. 1081-1095.

[15] Ronald A. Skoog, Thomas C. Banwell, Joel W. Gannett, Sarry F. Habbiby, Marcus Pang, Michael E, Rauch, and Paul Toliver. "Automatic Identification of Impairments Using Support Vector Machine Pattern Classification on Eye Diagrams", IEEE Photonics Technology Letters,Vol-18, No-22, 2006,pp. 2398 – 2400.

[16] Rui Pereira, and Luis Seabra Lopes, "Learning Visual Object Categories with Global Descriptors and Local Features", Proceedings of the 14th Portuguese Conference on Artificial Intelligence, Progress in Artificial Intelligence, Portugal. 2009, pp. 225 – 236.

[17] Shailedra Kumar Shrivastava, and Sanjay S. Gharde "Support Vector Machine for Handwritten Devanagari Numeral Recognition", International Journal of Computer Applications, Vol-7, No-11, 2010, pp. 9-14.

[18] Xiangjin Zeng, Xinhan Huang, and Min Wang. "Research of Invariant Moments and Improved Support Vector Machine in Micro-Targets Identification", Journal of Applied Science, Vol-8, No-21, 2008, pp. 3969-3974.

[19] Xin-Han Huang, Xiang-Jin Zeng, Min wang "SVM-Based Identification and Un-calibrated Visual Servoing for Micro-Manipulation", International Journal of Automation and Computing, Vol-7, No-1, 2010, pp.47-54.

[20] Zhenchun Lei, Yingchun Yang, and Zhaohui Wu., "Ensemble of Support Vector Machine for Text-Independedt Speaker Recognition", International Journal of Computer Science and Network Security, Vol- 6, No-5A, 2006, pp. 163 – 167.

R.Muralidharan is a Ph.D student at Anna University of Technology, Coimbatore, India doing research in the field of Object Recognition Systems. He received M.Sc. Computer Science from Bharathidasan University in 2001 and M.Phil from Bharathiyar University in 2005. He is presently working as Assistant Professor in Computer Applications, KSR College of Engineering, Tiruchengode, Namakkal, India. He is a life member of ISTE and IACSIT. His research interests are in Image Processing, Object Recognitions and Neural Networks.

Dr.C.Chandrasekar received the B.Sc degree and M.C.A degree. He received his PhD from Periyar University, Salem at 2006. He worked as Head of the Department, Department Of Computer Applications at KSR College of Engineering from 2007. He has been working as Associate Professor in the Department of Computer Science at Periyar University, Salem. His research interest includes Mobile computing, Networks, Image processing, Pattern Recognition and Data mining. He is a senior member of ISTE, CSI. He was a Research guide at various universities in India. He has been published more than 50 technical papers at various National/ International Conference and Journals

Neural networks for error detection and data aggregation in wireless sensor network

Saeid Bahanfar[1], Helia Kousha[2] and Ladan Darougaran[3]

[1] Department of Computer engineering ,Islamic Azad University, Science and Research Branch
Tabriz, IRAN

[2] Young Researchers Club ,Tabriz Branch ,Islamic Azad University
Tabriz, IRAN

[3] Young Researchers Club ,Tabriz Branch ,Islamic Azad University
Tabriz, IRAN

Abstract

Correct information and data aggregation are very important in wireless sensor networks because sending incorrect information by fault sensors make to wrong decision about environment and increasing defective sensor during the time incorrect data decries reliability of wireless sensor networks. Previous methods have Problems such as there are fault sensors in wireless sensor network therefore wrong data are sent to CH by these sensors.

In this paper apply the neural network within the sensors, fault sensors and wrong data are discovered and eliminated. That is increased efficiency and reliability and longevity sensor networks.

Keywords: *CH, sink, Base Station, neural networks, error detection, data aggregation, reliability*

1. Introduction

Neural networks are a new tool to analyze complex and difficult issues, new strategies must be introduced. Neural networks are computer algorithms based on stimulus and response structure of the human brain have been the model. These networks often learn to map input - output a set of templates and are used samples. Functional relationships between variables "learned" are defined without the need is the relationship between individual variables. Neural networks to solve the problems that the relationship between variables is not clear, it isn't very useful. [1,2,3]. Recent technological advances such as sensors, electronics, computing devices, causing researchers tendency towards wireless sensor networks. Wireless sensor network typically consists of a large number of inexpensive sensor nodes, multi-functional, and limited energy and computing capabilities and communication [7,8]. We distributed sensors in the environment and clustered with clustering techniques them and we have chosen a cluster head for each cluster. Sensors send their information to CH. It using the techniques of data aggregation, it sends result to the sink, which this schema saves energy. Many sensors may be produce repetitive data with data aggregation schemas can reduce extra data. Different schemas are explained in [4,5]. Considering processing consumption energy less than transmits, data aggregation is very important because that, many protocols has been used this technique. We can use signal processing techniques, one of them is Data Fusion obtained more accurate information. Faulty sensor sends wrong data to CH and it sends to base station, and caused false information to be processed this is disadvantage of previous methods.

In this paper we presented embedded neural network into sensors and it is trained when the placement within them, which increases the reliability of neural network. We already have been determined the environment and therefore we know initial information about the environment. Education Network have done with them so that when we detect error with new data and neural network's result, which we can improve data collection in sensor networks. Section 2 a summary of neural networks and Section 3 the idea of using neural network within the sensor and check the results in Section 4.

2. Related work

Some of the data aggregation schemas [9,10] network is clustered at first. Then CH applies data aggregation. Lotfinezhad and Liang in [11] have tried to consider the effect of data to some extent dependent on the efficiency of clustering methods in data aggregation. Kstryn and his friends [12] have tried to estimates data of a node in a certain time with a third degree equation and instead send their data, send the polynomial coefficients. Dasgupta and his colleagues [13] have assumed every node able to do data aggregation into network. They presented a method based on the assumption for increasing the lifetime of the network. Beaver and Sheref in [14] have proposed an algorithm that tries to route a group of similar sensors discover (which sensors that produce the same data). In [10] sensors send model of data that shows how all the sensor's data based on predetermined intervals are subject to change. [11] Using the above negotiation transmission of extra data can be removed. In [15] published directly have been proposed as data gathering protocol for sensor networks which aims to monitor events that usually using a small number of nodes [16] a query system network. Roddy and Jack in [14] have been used a machine learning techniques which sensor nodes only send special data to the sink. Learning algorithm used in this way, the sink performed and then the results will be published on the network. Liang and colleagues [17] have used Q-learner for each node because they send their data along the path whit maximum rate of aggregation. In [18] are presented a method for data aggregation using learning automata.
Where are neural nets begin used?
The study of neural network is an extremely interdisciplinary field, both in its development and in its application. A brief sampling of some of the areas in which neural networks are currently being applied suggests the breath of their applicability. The examples range from commercial successes to areas of active research the show promise for the future. For example applied neural networks is signal processing, control [Nguyen & Widrow, 1989; Miller, Sutton, & Werbos, 1990], pattern recognition [Le Cun al., 1990], medicine [Anderson, 1989; Andeson Golden, and Murphy, 1989], [HechtNilson 1990], speech production, speech recognition and in the our idea used in wireless sensor network (WSN).

Who is developing neural networks?
This section presents a very brief summary of this history of neural networks, in terms of the development of architectures and algorithm that are widely used today. Results of a primarily biological nature are not included, due to space constraints. They have however, served as the inspiration for a number of networks that are applicable to problems beyond the original ones studied. The history neural networks show the interplay among biological experimentation modeling, and computer simulation / hardware implementation. Thus the field is strongly interdisciplinary. The 1940s beginning of neural networks: Warren McCulloch and Walter Pitts designed what are generally regarded as the first neural networks. Then Donald Hebb, a psychologist at McGill University, designed the first learning law for artificial neural network in 1947.The 1950s and 1960s is first golden age of neural networks: algorithm today neural networks are often viewed as an alternative to traditional computing; it is interesting to note that John Van Neumann, "the father of modern computing," was keenly interested in modeling the brain. Johnson and Brown (1988) and Anderson and Rosenfeld (1988) discuss the interaction between von Neumann and early neural network research such as Warren McCulloch, and present further indication of van Neumann's view of the direction in which computers would develop.

3. Propose Idea

In this article we let embedded neural network into each sensor. Each sensor sense data from the environment and also which has sense of sensor data as well as its adjacent receives. So that each sensor to communicate with others without the knowing their location therefore sensors are not dependent on location, for example, if the effect of environment factors such as wind, sensors are moved and they identify new neighbors they haven't problem and then work with new neighbors. Figure (1) How to communicate the sensor has been shown that the sensor broadcast data to radius R and there are sensors in the radius R. If they need the data; they can receive data. The presented idea due to neural networks within the sensor; therefore sensors need data by the neighboring sensors. These data are input for neural networks using training neural networks.

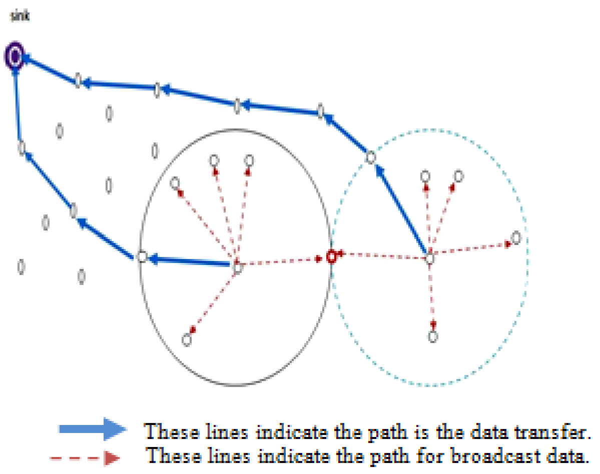

sink
These lines indicate the path is the data transfer.
These lines indicate the path for broadcast data.

Fig. 1 A way to receive information from nearby sensors

Function neural network train with the data; they received from neighbor sensors and then they produce a data (That is generated neural network's data). This data compare with sense data by the same sensor and then calculate rates error between the two data (production data with neural network's data) and call this error α.

$$|\text{Data neural network - Data sensor}| = \alpha. \qquad (1)$$

We consider β for accuracy of production data. If $\alpha > \beta$ then the sensor is defective (unsafe); therefore sensor and his data remove the wireless network (intercept to send unsafe data to CH, when happen this we save energy because we prevent consumption energy for sending unhealthy data to the CH). But if $\alpha < \beta$ then sensor is safe, therefore sensor can send data to CH (correct data is sent). Because the sensor adapt to changing environment and it's education is not only in the production stage, the neural network trains whit the own data and data received from neighboring sensors that are correct (according to above description) and we make update training of neural networks and we increase neural network's reliability. If we fail to discover and eliminate faulty sensors and the sensors can still send incorrect data to CH and this is continue whit increasing the number of faulty sensors to more false data will be sent to CH and because CH decide based on data received from sensors in cluster and CH dose data aggregation so CH can't decide and we can't trust the network's result after a certain time (This sensors of WSN haven't neural network), but we can trust the result of network which the sensors of WSN have neural network because increase reliability of WSN and lifetime of WSN.

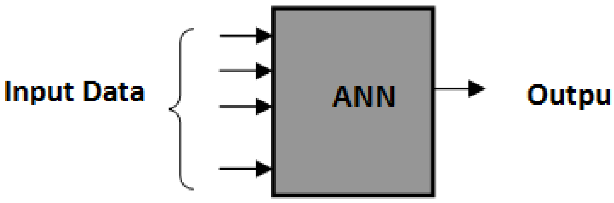


Fig. 2 neural network

How to schedule and arrange the input data to neural network. Figure (2) data of adjacent neighbors as input neural network time (t) and generated a output at the time (t_i) then compared this output whit the sense data by the sensor at the time(t_i+1), inputs at the time (t_i+2)are for training the neural network, this time training is applied with the supervisor then neural network is updating, at the time (t_i+3)all stages begin at first. Instructions sensor is as follows:

```
If sense data == true then

    Continue;
Else
     Turn off;
```

3.1 Sensor structure in our idea

Components of each sensor is shown in figure 3:

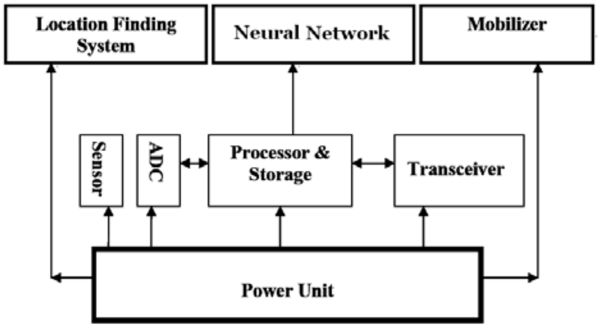


Fig. 3 Interior Sensor

Devising neural network within the sensors author to increase processing and because processing uses less energy than post processing so the amount of energy that we lose is more less than energy that we use to send data's and its reason of that we use neural network in the sensors and the other reason is it increases longevity of the network and increases reliability about the network.

Figure (4) a. It considered the normal case of sensor network and in this shape when the failure occurs the sensor don’t notice that and always send the incorrect data with the energy amount of NW joules that N is length and W is weight of the way during the course and this faulty sensor continuously sends this incorrect data that cause the increasing of the energy of the other sensors of the network.

Figure (4) B. Because we have neural network in the sensor at the first of processing diagnosed that the data is incorrect. This incorrect data uses some energy for processing but after this, this data will never send to base station. Although neural network is considered redundancy for the sensor network but this redundancy is effective in using of energy because as the instructions provided the faulty sensor eliminate and because it doesn’t send a fault data the energy can be saved. There is another way for the implementation neural network: We use nerve into the base station and this way not only decrease processing but also falsely stored the energy and also it includes some limits, for example it limited us that the sensor is be implant to know the location of the sensors and its neighbors and if a sensor shifts to other location it has trouble to identify its neighbors.

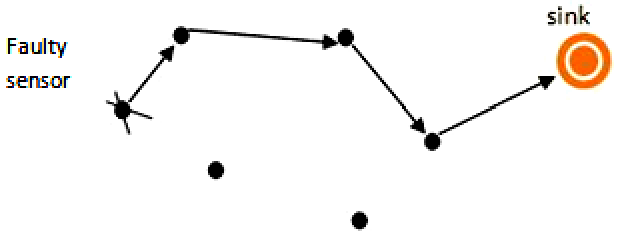


Fig. 4_a sensor without neural network

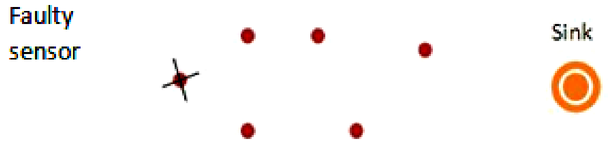


Fig. 4_b sensor with neural network

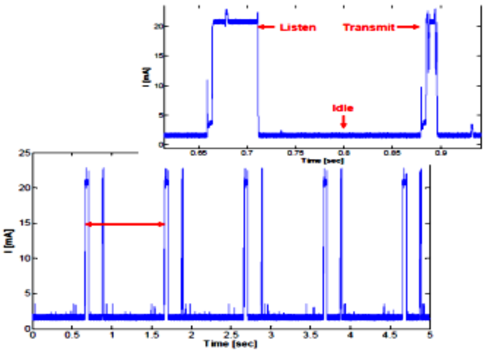


Fig. 5 consumption energy transferring data and processing data

As seen in the figure 5 transferring data uses more energy than processing data.

4. Simulation to verify the idea

4.1 Energy

If we assume that collecting data is time based, cluster collects the level of the data and the region is clustered but in the sensors there isn’t any neural network and after a certain time all sensors start to send data to CH if the data was correct or incorrect! Sending to cluster is been done but in our idea sensors can decide that the generated data is correct or not and after that it be sure that the data is correct and then it goes to send it.

As sending data from any sensor in neural to cluster uses some power when the sensor is broken it use this energy again and send an incorrect data to CH (a useless send that just uses the energy in the sensors network and also reduce the accuracy of decision in the CH).in our idea this amount of energy that broken sensor uses to send the fault data to CH is stored in the network.(These broken nodes are the source of energy in the network that they can use in the routing of other clusters data to sink, after that they use they energy to send unbroken sensor’s data to sink).

If in case that there isn’t any neural network in the sensors we have assumedly 100 sensors in a cluster each time we send data to CH all of the sensors attempt to send data to CH and always the certain amount of energy have to used (if the sensors produced data is correct or not)we will use 100x W power (if any sensor use x W power).. The diagram 6 shows the use of energy with increasing breakdown in the cluster that its sensors have neural network to 50 breakdown of 100 of the network with increasing the measure of breakdown the amount of energy using is like this:

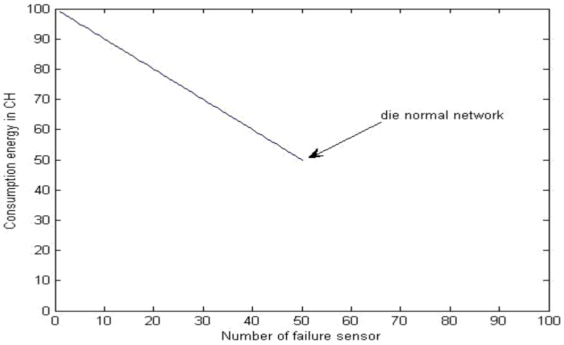


Fig. 6 energy consumptions

4.2 Data collection in networks without neural network

Currently available network sensitive sensors for collecting data work like this way that the sensed data with sensors send to the CH and CH collect the data's. If we consider some sensors in the environment with increasing of the sensor failure CH make a false decision about its area and if at least half of sensors work incorrectly the cluster dead will happen and limited cluster will send a false data to sink. The following chart shows the same. By this diagram (2) if we consider a cluster with 100 sensors (at first we consider that all of the sensors are work correctly) if our sensors are safe they build data 7, 8 or 9 and if they were rotten they build data 0,1,2,3,4,5,6, or 10. (That the data range in this environment is between 0 to 10) because sensors are distributed in clusters using a majority vote when the CH has healthy majority of sensors (with using of aggregating data's) the correct answer sends from CH to sink (the correct answer is 7, 8 or 9) but by this diagram when the most of sensors are defective the incorrect data is send to sink.

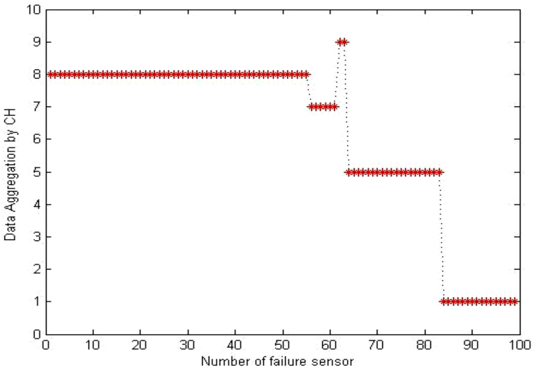


Fig. 7 Data aggregation networks with neural network

4.3 Data aggregation networks with neural network embedded within the sensor

In our idea to prove the accuracy of data produced in each sensor we embedded neural network in each of them. To simulate the above environment that we implement with normal sensors or sensors without neural network at this time we implement it with sensors that they have neural networks and by the below diagram with increasing of failure of sensors still data sent from CH to sink (data is just one of the data 7, 8 or 9 and the range of data is between 0 and10). When death has occurred in the regular network or more than half of the sensors are broken (means that the aggregate data is the result of incorrect data) by the below diagram this network still make correct data's and send it to sink.

Even with only 1 healthy sensor in the environment the network continue its life (it means death of the network don't be happen) the diagram tells it:

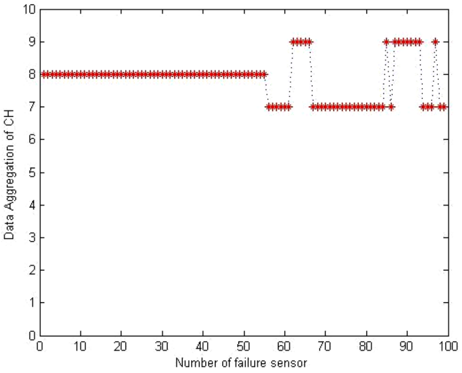


Fig. 8 Data aggregation

4.4 Increased longevity of Network

Bye the diagrams 2 and 3 because increasing the number of sensors failure still the network that its sensors have neural network give us the correct answer and it shows us that the network death don't happen with increasing of the corrupted sensors.

By the diagram 2 death of cluster (small sensor network) is happen if the 61 sensor of 100 sensors corrupted. But the sensor network that the sensors have neural network or data isn't send from sensor to CH or if data will be sent we can trust that data is correct.

5. Conclusion

Because each sensor depends on the performance of its neural network and neural network performance is the way that its performance increase at the time that you learn and in this passage studying neural network is continued and this work cause the increasing of ensure of neural network with the time going up and it cause increasing reliability of the wireless sensitive networks.

And neural network also can increase the longevity of wireless sensor network. Another desired result is achieved by using this method is identify and delete incorrect data element that they are destructive, redundancy that created by using neural network in the sensors is prevent transmission of false information and it results in energy consumption is being optimized.

References

[1] Krishanamachari B., D. [1] Krishanamachari B., D. Estrin, S. Estrin, S. Wicker “ ,Modeling Data Centric Routing inWireless Sensor Networks ,” Proc. Wicker, "Modeling Data Centric Routing inWireless Sensor Networks", Proc. of the IEEE INFOCOM, New York, USA, Jun2002 . of the IEEE INFOCOM, New York, USA, Jun. 2002

[2] Dr. [2] Dr. MBmenhaj “Artificial intelligence” R.Beal & T,Jackson "Neural Computing:An Introduction. ” MBmenhaj "Artificial intelligence" R. Beal & T, Jackson "Neural Computing: An Introduction".

[3] IFAkyildiz,w.su,Y.Sankarasubramaniam,E.Cayirei ”,wireless sensor networks:a survey ,” Computer networks ,(2002)38pp.393-422 [7] IFAkyildiz, w.su, Y. Sankarasubramaniam, E. Cayirei, "wireless sensor networks: a survey", Computer networks 38 (2002), pp.393-422

[4] R. Shah and J. Shah and J. Rabaey, Energy Aware Routing for Low Energy Ad Hoc Sensor Networks, in Rabaey, Energy Aware Routing for Low Energy Ad Hoc Sensor Networks, in

[5] Proceedings of the IEEE Wireless Communicationsand Networking Conference (WCNC), Orlando,Florida ,March 2002. Proceedings of the IEEE Wireless Communicationsand Networking Conference (WCNC), Orlando, Florida, March 2002.

[6] R. Virrankoski and A. Virrankoski and A. Savvides, TASC : Topology Adaptive Spatial Clustering for Sensor Networks, Second IEEE Intl. Savvides, TASC: Topology Adaptive Spatial Clustering for Sensor Networks, Second IEEE Intl. Conf. Conf. on Mobile Ad Hoc and Sensor systems", Washington, DC ,November, 2005. on Mobile Ad Hoc and Sensor systems ", Washington, DC, November, 2005.

[7] S. Soro and W. Soro and W. Heinzelman, Prolonging the Lifetime of Wireless Sensor Networks via Unequal Clustering, Proceedings of the 5th International Workshop on Algorithms for Wireless, Mobile, Ad Hoc and Sensor Networks (IEEE WMAN '05), April 2005. Heinzelman, Prolonging the Lifetime of Wireless Sensor Networks via Unequal Clustering, Proceedings of the 5th International Workshop on Algorithms for Wireless, Mobile, Ad Hoc and Sensor Networks (IEEE WMAN '05), April 2005.

[8] M. Lotfinezhad and B. Liang, Effect of partially correlated data on clustering in wireless sensor networks , in Proceedings of the IEEE International Conference on Sensor and Ad hoc Communications and Networks) SECON), Santa Clara, California, 2004.. Lotfinezhad and B. Liang, Effect of partially correlated data on clustering in wireless sensor networks, in Proceedings of the IEEE International Conference on Sensor and Ad hoc Communications and Networks (SECON), Santa Clara, California, 2004.

[9] C. Guestrin, P. Guestrin, P. Bodik ,R. Bodik, R. Thibaux, M. Thibaux, M. Paskin and S. Paskin and S. Madden, Distributed Regression : An Efficient Framework for Modeling Sensor Network Data, 2004. Madden, Distributed Regression: An Efficient Framework for Modeling Sensor Network Data, 2004.

[10] K. Dasgupta, K. Dasgupta, K. Kalpakis and P. Kalpakis and P. Namjoshi, An Efficient Clustering-based Heuristic for Data Gathering and Aggregation in Sensor Networks, IEEE Wireless Communications Conference, Vol.4, No. Namjoshi, An Efficient Clustering-based Heuristic for Data Gathering and Aggregation in Sensor Networks, IEEE Wireless Communications Conference, Vol.4, No. .2003 ,1 1, 2003.

[11] Fausett,L., “Fundamentals of Neural Networks".1994. [3] Fausett, L., "Fundamentals of Neural Networks" .1994.

[12] Stephen Wicker “ The Effect of Imperfect Error Detection on Reliability Assessment via Life Testing ” IEEE TRANSACTION ON SOFTWER ENGENEERING.VOL.20.NO .2,FEBRUARY1994 Stephen Wicker "The Effect of Imperfect Error Detection on Reliability Assessment via Life Testing" IEEE TRANSACTION ON SOFTWER ENGENEERING.VOL.20.NO .2, FEBRUARY 1994

[13] Deborah Estrin “ Dissociable Executive Functions in the Dynamic Control of Behavior:Inhibition, Error Detection, and Correction.” [5] Deborah Estrin "Dissociable Executive Functions in the Dynamic Control of Behavior: Inhibition, Error Detection, and Correction". NeuroImage–1820 ,17 (2002) 1829doi:10.1006/nimg.2002.1326 NeuroImage 17, 1820-1829 (2002) doi: 10.1006/nimg.2002.1326

[14] Krishanamachari B., D. [6] Krishanamachari B., D. Estrin and S. Estrin and S. Wicker “ ,The Impact of Data Aggregation in Wireless Sensor Networks ,” Proc. Wicker, "The Impact of Data Aggregation in Wireless Sensor Networks", Proc. of the International Workshop of of the International Workshop of Distributed Event Based Systems (DEBS), Vienna, Austria, Jul. 2002, pp. Distributed Event Based Systems (DEBS), Vienna, Austria, Jul. 2002, pp. -575 ..578 575-578

[15] Y. Xu, W. Xu, W. C. C. Lee, J. Xu, and G. Lee, J. Xu, and G. Mitchell ,Processing Window Queries in Wireless Sensor Networks, IEEE International Conference on Data Engineering (ICDE'06 ,(Atlanta, GA, April 2006. Mitchell, Processing Window Queries in Wireless Sensor Networks, IEEE International Conference on Data Engineering (ICDE'06), Atlanta, GA, April 2006.

[16]R. Rosemark and WC Lee, Decentralizing Query Processing in Sensor Networks, the Second International Conference on Mobile and Ubiquitous Systems: Networking and Services (Mobiquitous'05), San Diego, CA, July, 2005, pp. 270-280.

[17] B. Karp and H. K ung, Greedy perimeter stateless routing for wireless networks , In Proceedings of the Sixth Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom 2000), pages 243–254, Boston, MA, August 2000.

[18] P. Beyens, M. Peeters, K. Steenhaut and A. Nowe, Routing with Compression in Wireless Sensor Networks: a Q-learning Approach , In "Fifth European Workshop on Adaptive

Agents and Multi-Agent Systems (AAMAS 05), Paris, France.", 2005.

Saeid Bahanfar received the B.Sc. degree in Computer Software Engineering from Payam Noor University (PNU), Tabriz branch, Iran in 2008._ Currently, he is a M.Sc. student of Computer System Architecture in Islamic Azad University, Tabriz branch, Iran. His research interests include Residue Number System and VLSI Design, wireless sensor network, Neural network.

Helya Kousha received her B.Sc. in Computer Software Engineering from Islamic Azad University, Shabestar branch, Iran in 2008. Currently, she is a M.Sc. student of Computer System Architecture in Islamic Azad University, Tabriz branch, Iran. Her main research interests include Computer Arithmetic, Residue Number System, wireless sensor network.

Ladan Darouagarn was born in Tabriz, Iran, on May 29, 1983. She received the B.Sc. degrees from University of Shabestar (Shabestar, Iran) and M.S.E. student in Islamic Azad University, Tabriz Branch in 2011. Her research interests are in the data aggregation in wireless sensor network. She is a member of Young Researchers Club.

Automatic image clustering using a swarm intelligence approach

Salima Ouadfel[1], Mohamed Batouche[2] and Abdlemalik Ahmed-Taleb[3]

[1] University of Batna Computer Science department
Batna 33000, Algeria

[2] COEIA – CCIS, King Saud University,
Riyadh, Kingdom of Saudi Arabia

[3] LAMIH UMR CNRS UVHC
8530Valenciennes, France

Abstract

In order to implement clustering under the condition that the number of clusters is not known a priori, we propose in this paper ACPSO a novel automatic image clustering algorithm based on particle swarm optimization algorithm. ACPSO can partition image into compact and well separated clusters without any knowledge on the real number of clusters. ACPSO used a novel representation scheme for the search variables in order to determine the optimal number of clusters. The partition of each particle of the swarm evolves using evolving operators which aim to reduce dynamically the number of clusters centers. Experimental results on real images demonstrate the effectiveness of the proposed approach.

Keywords: *Image clustering, swarm intelligence, Particle swarm optimization, automatic clustering.*

1. Introduction

Image segmentation is an important technology for image processing, and also is a fundamental process in many image, video, and computer vision applications. The goal of image segmentation is to cluster pixels into salient image regions, such as regions corresponding to individual surfaces, objects, or natural parts of [1]. Clustering process aims to partition the image into clusters such that the pixels within a cluster are as homogenous as possible whereas the clusters among each other are as heterogeneous as possible with respect to some similarity measure.

Several clustering methods are provided in the literature [2]. They fall into two categories: hierarchical and partitioning methods. Hierarchical methods proceed by stages producing a sequence of partitions, where each partition corresponds to a different number of clusters. A hierarchical algorithm yields a tree representing the nested grouping of patterns. Partitioning methods obtain a single partition of the pixels by moving pixels iteratively from one group to another, starting from an initial partition. An extensive survey of various clustering techniques can be found in [2]. The focus of this paper is on the partitional clustering algorithms.

Hard or crisp partitional clustering [3] and fuzzy partitional clustering [4] are two partitioning clustering algorithms such that hard clustering assigns each data point to only one cluster while fuzzy clustering assigns each data point to several clusters with varying degrees of memberships. The most widely used hard partitioning algorithm is the iterative K-means approach [5, 6, 7]. In the K-means algorithm, pixels with similar features like gray levels or colors are grouped in the same cluster. The clustering is obtained by iteratively minimizing a cost function that is dependent on the distance of the pixels to the cluster centers. The major problem with this algorithm like most of the existing clustering algorithms is that its result is sensitive to the selection of the initial partition, it may converge to local optima and it requires the a priori specification of the number of clusters *K*.

To deal with the limitations existing in the traditional partition clustering methods, a number of new clustering algorithms have been proposed with the inspiration coming from observations of natural processes [8].

In order to remedy the drawbacks of K-means, this paper proposes a new automatic image clustering algorithm based on a modified version of particle swarm optimization. The proposed algorithm, called by us the ACPSO (Automatic Clustering with PSO) effectively search for both the optimal cluster centers positions and the number of effective clusters, and this with minimal user interference. ACPSO has the following characteristics: (1) particles can contain different cluster number in a range defined by minimum and maximum cluster number, (2) Particles are initialized randomly to process different cluster numbers in a specified range, (3) The goal of each particle is to search the optimum number of clusters and the optimum cluster centers, (4) Three new evolving operators are introduced to evolve dynamically the partitions encoded in the particles.

The paper is organized as follows. Section 2 defines the clustering problem in a formal language and gives a brief overview of a previous works done in the field of unsupervised partitional clustering. Section 3 presents a description of PSO algorithm. Section 4 outlines the proposed ACPSO algorithm. In section 5, we present the experimental results as well as a comparative study. Finally, conclusion is drawn.

2. Scientific background

2.1 Problem definition

The clustering problem can be formally defined as follows. Given a data set $Z = \{z_1, z_2, ...z_n\}$ where z_i is a data item and n is the number of data items in *Z*. The clustering aims to partitioning *Z* into *K* compacts and well separated clusters.

Compactness means that members of a cluster are all similar and close together. One measure of compactness of a cluster is the average distance of the cluster instances compared to the cluster center.

$$compactness(c_j) = \frac{1}{n_j} \sum_{z_i \in C_j} \left(z_i - m_j\right)^2 \qquad (1)$$

where m_j is the center of the *j*th cluster c_j and n_j is its cardinal. Lower value of *compactness* (c_j) is better.

Thus, the overall compactness of a particular grouping of *K* clusters is just the sum of the compactness of the individual clusters

$$compactness = \frac{1}{n} \sum_{j=1}^{K} \sum_{z_i \in C_j} \left(z_i - m_j\right)^2 \qquad (2)$$

Separability means that members of one cluster are sufficiently different from members of another cluster (cluster dissimilarity). One measure of the separability of two clusters c_i and c_j is their squared distance.

$$separability(c_i, c_j) = \left\| m_i - m_j \right\| \qquad (3)$$

where m_i and m_j are the center of the *i*th and *j*th cluster respectively.

The separability of the partition of *K* clusters could be defined as following:

$$Separability = \sum_{i=1}^{K} \min_j separability(c_i, c_j) \qquad (4)$$

The bigger the distance, the better the separability, so we would like to find groupings where separability is maximized.

2.1 Unsupervised clustering algorithms

Clustering can be formally considered as a particular kind of NP-hard grouping problem [9]. This assumption has stimulated much research and use of efficient approximation algorithms.

One of the most frequently used clustering algorithms is the iterative K-means algorithm [10, 11]. The K-means algorithm starts with *K* cluster centers randomly selected using some heuristics. Each data item in the data set is then assigned to the closest cluster center according to a distance measure. The centers are updated by using the mean of the associated items. The process is repeated until some stopping criterion is verified. Although the k-means algorithm has been widely used due to its easy implementation, it has two major drawbacks: it is too sensitive to the initial clusters centers and it needs to specify the number of clusters in advance. However, in many practical cases, it is impossible to determine the exact cluster number in advance. Under these circumstances, the k-means algorithm often leads to a poor clustering performance.

In the literature, many approaches to finding dynamically the number of clusters has been proposed. In [12], the ISODATA (Iterative Self-Organizing data Analysis technique) was proposed. Like the K-means algorithm, ISODATA assigns each item to the closest cluster center; however, it adds division of a cluster c_i into two clusters if the cluster standard deviation of c_i exceeds a user-specified threshold th_{div}, and processing of fusion of two clusters if the distance between their centers is smaller than another user-specified threshold th_{merg}. Using this variant, the optimal partition starting from any arbitrary initial partition can be obtained. However, it requires many parameters to be specified by the user. In [13], the authors proposed SYNERACT, which combines K-means with hierarchical descending approaches. In [14] Rosenberger and Chehdi introduced a new improvement to the K-means algorithm. During each step of the clustering process, from a set of K clusters, a cluster with the higher intra_cluster distance is chosen for splitting into two clusters. Next, the K-means algorithm is applied to the (K+1) clusters. The iterative procedure is repeated until a valid partition of the data items is obtained. Pelleg and Moore [15] proposed X-means algorithm which is based on the classical K-means algorithm with the model selection. Hamerly [16] proposed G-means algorithm which splits clusters that not fit a Gaussian distribution.

Since the problem of data clustering can be easily viewed as a complex optimization problem [17], several optimization algorithms have been used for optimizing the cost function and to find the optimal number of clusters. For example, in [18], the authors proposed a nonparametric Variable string length genetic algorithm (VGA), with real encoding of the cluster centers in the chromosome. In [19] a novel variable length GA (VLIGA) algorithm which is an improvement version of VGA was proposed with a modified mutation function. In [20], authors proposed an evolutionary-fuzzy clustering algorithm for automatically grouping the pixels of an image into different homogeneous regions. The algorithm does not require a prior knowledge of the number of clusters. The fuzzy clustering task in the intensity space of an image is formulated as an optimization problem. An improved variant of the differential evolution (DE) algorithm has been used to determine the number of naturally occurring clusters in the image as well as to refine the cluster centers. Bandyopadhyay proposed in [21] a Variable String Length Simulated Annealing (VFC-SA) algorithm, which applied a simulated annealing algorithm to the fuzzy c-means clustering technique and used a cluster validity index measure as the energy function. Tseng and Yang [22] proposed a genetic algorithm based approach for the clustering problem. The proposed method can search for a proper number of clusters and classify non overlapping objects into these clusters. Lin et al. [23] presented an automatic genetic clustering algorithm based on a binary chromosome representation. Lai [24] adopted the hierarchical genetic algorithm to solve the clustering problem. In the proposed method, the chromosome consists of two types of genes, control genes and parametric genes. The control genes are coded as binary digits. The parametric genes are coded as real numbers to represent the coordinates of the cluster centers. The total number of “1” represents the number of clusters. In [25] authors proposed an algorithm to determine the optimal number of clusters by applying SA to cluster microarray data. In their method, first the fuzzy k-means algorithm is used to minimize the sum of within-cluster distance, then, the optimal number of clusters is obtained from the SA algorithm. In [26], authors proposed a dynamic clustering algorithm based on a modified version of classical Particle Swarm Optimization (PSO) algorithm, known as the Multi-Elitist PSO (MEPSO) model. A new particle representation scheme has been adopted for selecting the optimal number of clusters from several possible choices. It also employs a kernel-induced similarity measure instead of the conventional sum-of-squares distance. In [27] a new fuzzy clustering algorithm is proposed by combining the possibility clustering and ISODATA clustering algorithm. This new algorithm not only can determine the number of clusters dynamically with the degree of possibility of each date point, but also can reduce the number of input parameters of ISODATA algorithm. In [28] an approach for solving the automatic clustering of the Gene Ontology is proposed by incorporating cohesion-and-coupling metric into a hybrid algorithm consisting of a genetic algorithm and a split-and-merge algorithm. In [29] authors address the problem of cluster number selection by using a k-means approach that exploits local changes of internal validity indices to split or merge clusters. The split and merge k-means issues criterion functions to select clusters to be split or merged and fitness assessments on cluster structure changes. In [30] authors propose a Bacterial Evolutionary clustering algorithm, which can partition a given dataset automatically into the optimal number of groups. Experiments were done with several synthetic as well as real life data sets including a remote sensing satellite image data. The results establish the superiority of the proposed approach in terms of final accuracy. In [31] Omran et al. presented dynamic clustering PSO (DCPSO), which is, in fact, a hybrid clustering algorithm where binary PSO is used to determine the number of clusters while the traditional K-means method performs the clustering operation with this number of clusters. In [32], Abraham et al. combined the Fuzzy clustering algorithm with the multielitist PSO (MEPSO) to find automatically the number of clusters. In [33] authors proposed an evolutionary particle swarm optimization for data clustering. The proposed algorithm is based on the evolution of swarm generations. After each generation, the swarm dynamically adjusts itself in order to reach optimal position.

3. Particle swarm optimization

Particle swarm optimization (PSO) is a population-based evolutionary computation method first proposed by Kennedy and Eberhart [34]. It originated from the computer simulation of the individuals in a bird flock or fish school, which basically show a natural behavior when they search for some target (e.g., food). The PSO algorithm is initialized with a swarm of n particles randomly distributed over the search area with a random velocity and a random position. Each particle encodes a potential solution to the optimization problem. Particles flies through the search space and aims to converge to the global optimum of a function attached to the problem.

Each particle x_i in the swarm is represented by the following characteristics: the current position of the particle (p_i) and the current velocity of the particle (v_i). Its movement through the search space is influenced dynamically according to its personal best position *Pbest*, which is the best solution that it has so far achieved and its neighbors’ best position P_g. At each iteration t, the particle's new position and its velocity are updated as follows:

$$p_i(t) = p_i(t-1) + v_i(t) \quad (5)$$

$$v_i(t) = wv_i(t-1) + c_1 \times rand_1(p_{best} - p_i(t-1)) + c_2 \times rand_2(p_g - p_i(t-1)) \quad (6)$$

The parameter *w* is an inertia weight and it is equivalent to a temperature schedule in the simulated annealing algorithm and controls the influence of the previous velocity: a large value of *w* favors exploration, while a small value of *w* favors exploitation [35]. As originally introduced, w decreases linearly during the run from w_{min} to w_{max}. c_1 and c_2 are two constants which control the influence of the social and cognitive components such that $c_1 + c_2 = 4$. $rand_1$ and $rand_2$ are random values in the range [0,1].

Two topologies of neighborhoods exist in the literature: the *gbest* model and the *lbest* model. The *gbest* model maintains only a single best solution, called the global best particle, across all the particles in the swarm. This particle acts as an attractor, pulling all the particles towards it. The *gbest* offers a faster rate of convergence at the expense of robustness. The *lbest* model tries to prevent premature convergence by maintaining multiple attractors. In fact, *gbest* model is actually a special case of the *lbest* model. Experiments have shown that *lbest* algorithm converges somewhat more slowly than the *gbest* version, but it is less likely to become trapped in an inferior local minimum.

4. ACPSO algorithm

In this section, we describe an automatic image clustering algorithm based on a new version of particle swarm optimization algorithm, called ACPSO.

Let $Z = \{z_1, z_2, ...z_n\}$ be the image with *n* number of pixels. The ACPSO maintains a swarm of particles, where each particle represents a potential solution to the clustering problem. Each particle encodes an entire partition of the image Z. ACPSO tries to find an optimal partition $C = \{c_1, c_2, ...c_k\}$ of *K* optimal number of compactness and well separated clusters. In ACPSO, both the numbers of clusters as well as the appropriate clustering of the data are evolved simultaneously using the search capability of particle swarm optimization algorithm.

4.1 Particle representation

The initial population $P = \{X_1, X_2, X_3, ...X_{pop\_size}\}$ is made up of *pop_size* possible particles (solutions). For a user-defined maximum cluster number $K_{max,}$, a single particle x_i is a vector of K_{max} binary numbers 0 and 1 (flags) and K_{max} real numbers that represents the K_{max} cluster centers.

For a particle x_i, each probably cluster center m_{ij} (j=1…$K_{\max}$) is associated with a binary flag $\gamma_{ij}(j = 1....K_{\max})$. The cluster center m_{ij} is valid and so selected to clustering the image pixels, if it's corresponding flag $\gamma_{ij} = 1$ and invalid if $\gamma_{ij} = 0$. The total number of "1" implicitly represents the number of clusters encoded in a particle.

If due to the update of the position of a particle some flags in a particle exceed 1, it is fixed to 1 or zero, respectively. However, if it is found that no flag could be set to one in a particle (all cluster centers are invalid and so no selected), two random flags are selected and we re-initialize them to 1. Thus the minimum number of possible clusters is always 2.

Two examples of the particle structure in the proposed approach are shown in Figure 1.

$$[\ \overbrace{0,0,1,1,0}^{actvation-clusters-part}\ ,\ \overbrace{12.5,\ 45.7,\ 36.5,\ 22.5,\ 66.3}^{cluster-centers-part}]$$

Particle *i* represents 2 clusters, and the associated cluster centers are 36.5 and 22.5.Cluster centers 12.5, 45.7 and 66.3 are invalid and not used to clustering the image.

$$[\ \overbrace{1,0,1,0,1}^{actvation-clusters-part}\ ,\ \overbrace{39.5,\ 45.7,\ 26.5,\ 40.3,\ 33.3}^{cluster-centroids-part}]$$

Particle *i* represents 3 clusters, and the associated cluster centers are 39.5, 26.5 and 33.3. Cluster centers 45.7, 26.5 and 40.3 are invalid.

Figure1. Two examples of the particle structure in the ACPSO algorithm.

4.2 Population initialization

To generate the initial population of particles, we use in this paper the random generation strategy until all particles in a population are created. For a particular particle x_i , K_i cluster centers are randomly selected points from the given data set and K_i flags are randomly generated. Note that if the number of valid centers contained in a particle is less than two, then its flags are reinitialized.

4.3 Fitness evaluation

The fitness of a particle indicates the degree of goodness of the solution it represents. In this work, the fitness function of a particle is based on the Ray and Turi's validity criterion [36] proposed to color image segmentation using the $\frac{\text{int}\,ra}{\text{int}\,er}$ ratio with a multiplier function to avoid the selection of low cluster numbers. The criterion is defined as:

$$V(K) = \left(c \times N(2{,}1) + 1\right) \times \frac{\text{int}\,ra}{\text{int}\,er} \qquad (7)$$

where c=25 is a constant multiplier, K is the number of clusters found by the clustering algorithm and *N(2,1)* is a Gaussian function with mean 2 and standard deviation of 1. The intra and inter cluster distances represent respectively the compactness and the separability measures of clusters and are defined by Eq. (2) and Eq. (4).

A lower value of *V(K)* indicates a better quality of the clustering.

The Ray and Turi's measure based fitness function (to be maximized) for the particle x_i encoding K_i clusters is given by:

$$Fitness_i = \frac{1}{V(K_i) + eps} \qquad (8)$$

where *eps* is a small bias term equal to $2\text{x}10^{-4}$ and prevents the denominator of Eq. (8) from being equal to zero. When the algorithm converges, the particle that has the maximum Fitness value will be the optimal particle.

4.4 Evolving operators

The evolving operators are specifically designed to allow the number of the clusters of the particles to be changed dynamically. In the following, we describe each evolving operator.

4.4.1 Perturb operator

A valid cluster c_{ij} of the configuration encoded in a particle x_i is chosen randomly to be perturbed. The centre m_{ij} of the selected cluster c_{ij} is then modified as follows:

$$m_{ij}^{new} = m_{ij}^{old} + \delta * m_{ij}^{old} \qquad (9)$$

where m_{ij}^{new} and m_{ij}^{old} represent the new and the old cluster centre of the cluster c_{ij} . δ is a random number between [-1, 1].

Thus the cluster encoded by the particle is reconfigured, although the number of clusters belonging to it remains unaltered.

4.4.2 Split operator

For a particle x_i, we compute the compactness measure for each valid cluster according to Eq. (1). Let S the set of clusters c_{ij} $(j = 1...k_i)$ with the compactness measure higher than a threshold th_{split}. The threshold th_{split} is defined as the global compactness measure (see Eq. (2)) divided by the number of clusters of the particle x_i .

A cluster c_{ij} from the set S is selected for splitting into two new valid clusters, with the probability P_{split} defined as follows:

$$P_{split}(c_{ij}) = \frac{\frac{1}{n_j} \sum_{z_k \in c_{ij}} \left\| z_k - m_{ij} \right\|^2}{\sum_{s=1}^{K_i} \frac{1}{n_s} \sum_{z_k \in c_{is}} \left\| z_k - m_{is} \right\|^2} \qquad (10)$$

That is, the *sparser* cluster c_{ij}, the more possibly it is selected as the cluster for the split operator and vice versa.

The resulting number of clusters is K_i +1 and must be lower than K_{max}, otherwise, the split operator terminates.

4.4.3 Merge operator

For a particle x_i, first the pairwise separation distances D_{jl} between all distinct pairs of valid clusters (c_{ij},c_{il}) are calculated according to Eq. (3). Let S the set of pairs of valid cluster with the distance D_{jl} lower than a threshold th_{merge} . The threshold th_{merge} is defined as the average distance of D_{jl} for all distinct pairs of valid clusters.

A pair of distinct clusters (c_{ij},c_{il}) of S is selected for the merge operator with the probability P_{merge} defined as follows:

$$P_{merge}(c_{ij}, c_{il}) = 1 - \frac{D_{jl}}{\max(D_{jl})} \qquad (11)$$

where $\max(D_{jl})$ is the maximum pairwise separation distance between all distinct pairs of valid cluster centers from the set *S*.

The final number of clusters must be greater than 2, otherwise, the merge operator terminates.

Any one of the above-described evolving operators is applied for a particle if it is selected. The particle is selected with an adaptive probability P_e as in [37]. Let *gbest* the global best fitness of the current iteration; $\overline{Pbest}$ be the average fitness value of the population and $Pbest_i$ be the fitness value of the solution (particle) to be evolved. The expression for probability, P_e is given below:

$$P_e = \begin{cases} k_2 \times \dfrac{(gbest - Pbest_i)}{(gbest - \overline{Pbest})} & \text{if } Pbest_i > \overline{Pbest} \\ \mathrm{k}_4 & \text{if } Pbest_i \leq \overline{Pbest} \end{cases}$$

Here, values of k_2 and k_4 are kept equal to 0.5 [37]. This adaptive probability helps PSO to avoid getting stuck at local optimum.

The value of P_e increases when the fitness of the particle is quite poor. In contrast when the fitness of the particle is a good solution, P_e will be low so as to reduce the likelihood of disrupting good solution by evolving operators.

The framework of the ACPSO algorithm is given as follows:

```
1. Initialize the maximum cluster number $K_{max}$ and all the constant parameters;
2.Initialize each particle $x_i$ with random $k_i \in \{2,3,..,K_{\max}\}$ , randomly selected cluster centers, flags and initial velocities.
3. Initialize for each particle $x_i$ the $Pbest_i$
4. Initialize the gbest
5. For each particle $x_i$
      Calculate the fitness value $Fitness_i$ using Eq. 7.
      Set $Pbest_i = Fitness_i$.
      If ($\mathrm{Pbest}_i > gbest$) then set $gbest = Pbest_i$.
6. Update the position and the velocity of each particle according to Eqs. (5) and (6)
7. Apply randomly the evolving operators to alter the clusters centers of each particle
8. If termination criterion is satisfied go to step 9 else go to step 5
9. Segment the image using the optimal number of clusters and the optimal clusters centers given by the best global particle.
```

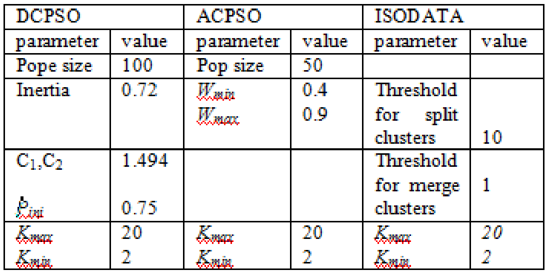
5. Experimental results

In order to evaluate the ability of our algorithm ACPSO to find the optimal clusters, we have tested it using natural images with varying range of complexity.

The performance of three dynamic clustering algorithms, ACPSO, DCPSO and ISODATA, were compared.

The parameter settings of DCPSO and ISODATA algorithm were determined by both referring to original papers and performing empirical studies. In Table 1, we report an optimal set-up of the parameters that gives the best results.

Table1. Parameter setup of the clustering algorithms for the image segmentation problem



DCPSO		ACPSO		ISODATA	
parameter	value	parameter	value	parameter	value
Pope size	100	Pop size	50		
Inertia	0.72	W_{min} W_{max}	0.4 0.9	Threshold for split clusters	10
C_1,C_2 p_{ini}	1.494 0.75			Threshold for merge clusters	1
K_{max} K_{min}	20 2	K_{max} K_{min}	20 2	K_{max} K_{min}	20 2

The clustering algorithms used in the experimental tests have been run several times for each test image. The optimal number of clusters has not been provided to any of the three optimization algorithm. Table 2 and Table 3 report the experimental results obtained over the grayscale images in terms of the mean and standard deviations of the number of clusters found and the final Turi measure reached by the three clustering algorithms. The results have been stated over 40 independent runs in each case.

Table 2. Number of clusters found by the clustering algorithms for real grayscale images.

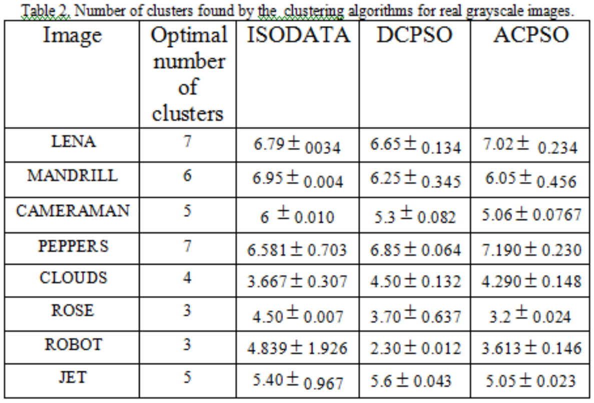


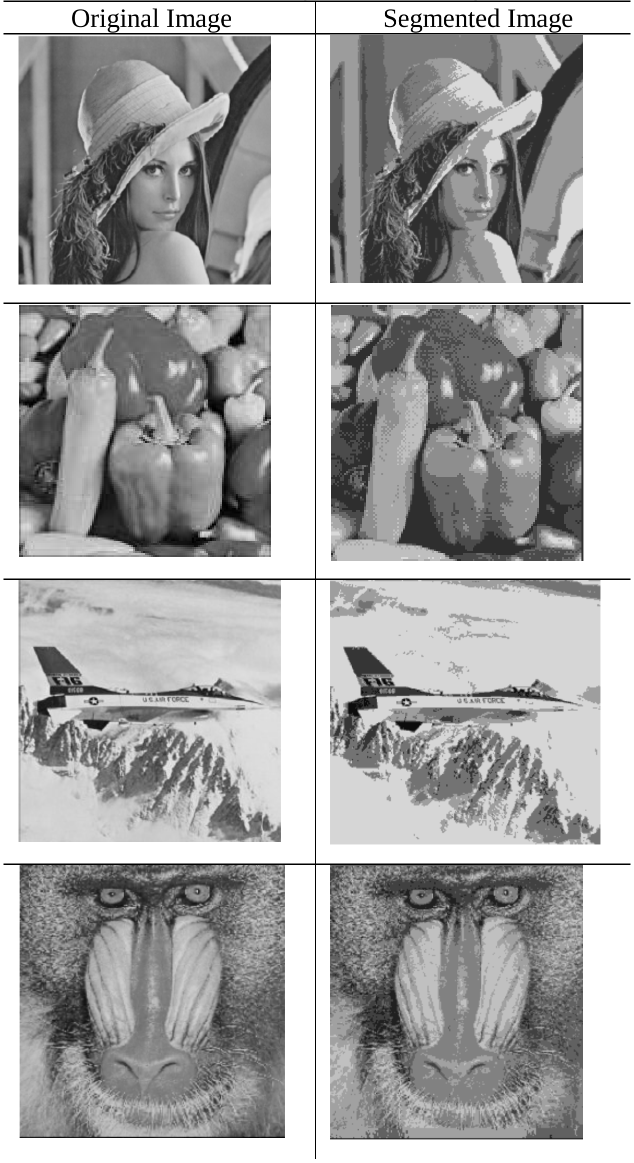
Image	Optimal number of clusters	ISODATA	DCPSO	ACPSO
LENA	7	6.79 ± 0034	6.65 ± 0.134	7.02 ± 0.234
MANDRILL	6	6.95 ± 0.004	6.25 ± 0.345	6.05 ± 0.456
CAMERAMAN	5	6 ± 0.010	5.3 ± 0.082	5.06 ± 0.0767
PEPPERS	7	6.581 ± 0.703	6.85 ± 0.064	7.190 ± 0.230
CLOUDS	4	3.667 ± 0.307	4.50 ± 0.132	4.290 ± 0.148
ROSE	3	4.50 ± 0.007	3.70 ± 0.637	3.2 ± 0.024
ROBOT	3	4.839 ± 1.926	2.30 ± 0.012	3.613 ± 0.146
JET	5	5.40 ± 0.967	5.6 ± 0.043	5.05 ± 0.023

Table 3. Automatic clustering result over real grayscale images using the Turi based fitness function over 40 independent runs.

	$(\text{Turi index})^{-1}$		
Image	ISODATA	DCPSO	ACPSO
LENA	0.19	0.16	**0.12**
MANDRILL	0.14	0.12	**0.10**
CAMERAMAN	0.097	0.089	**0.086**
PEPPERS	0.16	0.12	**0.10**
CLOUDS	0.094	0.074	**0.070**
ROSE	0.107	0.097	**0.084**
ROBOT	0.19	0.067	**0.052**
JET	0.098	0.070	**0.057**

From Tables 2-3 we can see that the proposed algorithm ACPSO outperforms the state of-the-art DCPSO and ISODATA algorithms for the present images related problems. The proposed algorithm is able to find the optimal number of clusters with better clustering result in term of the Turi cluster validity index.

Figure 2 shows the original images and their segmented counterparts obtained using the ACPSO algorithm.



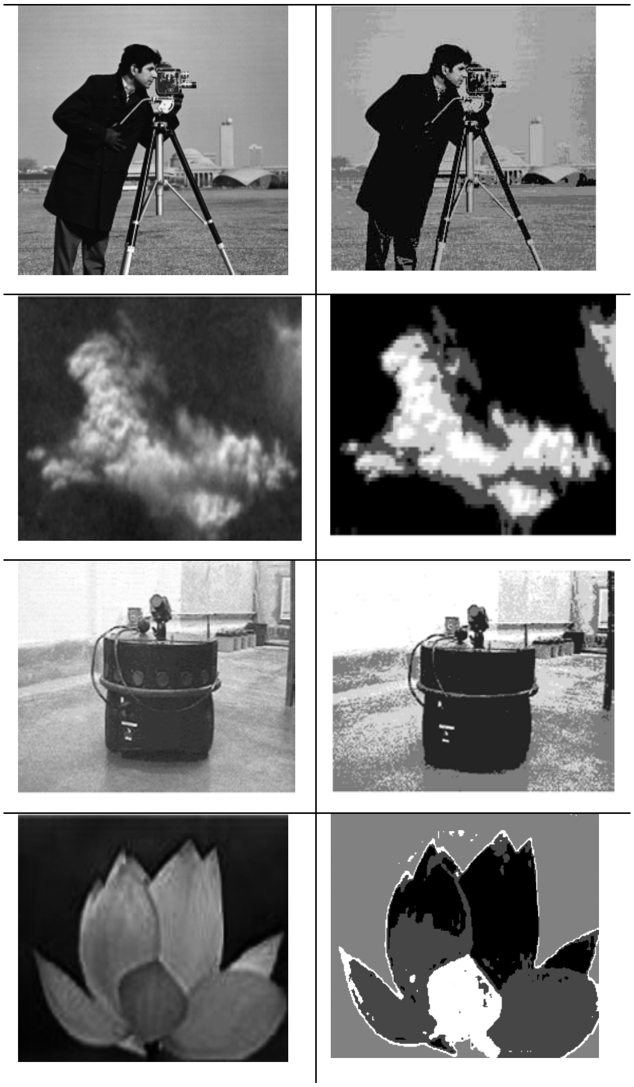


Figure 2. Samples of segmented images resulting from ACPSO

6. Conclusion

In this paper we have presented a new particle swarm optimization based method for automatic image clustering. ACPSO, in contrast to most of the existing clustering techniques, requires no prior knowledge of the data to be classified. ACPSO used a novel representation scheme for the search variables in order to determine the optimal number of clusters. Each particle encoded a partition of the image with a number of clusters chosen randomly from the set of the maximum number of clusters. The partition of each particle of the swarm evolves using evolving operators which aim to reduce dynamically the number of clusters centers. Superiority of the new method has been demonstrated by comparing it with ISODATA algorithm and a recently developed partitional clustering technique based on Particle Swarm Optimization (PSO) algorithm.

References

[1] Monga O. and Wrobel B. Segmentation d'images: vers une méthodologie, Traitement du Signal, 1987, 4(3), 169-193.

[2] Jain, A.K., Murty, M.N., Flynn, P.J. :Data clustering: a review. ACM Computing Surveys 1999, 31(3), 264–323.

[3] Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis. John Wiley and Sons, ManChichester (1973)

[4] Bezdek, J.C., Keller, J., Krishnampuram, R., Pal, N.R. Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. Kluwer Academic Publishers, Dordercht (1999)

[5] Forgy, E.W. Cluster Analysis of Multivariate Data: Efficiency Versus Interpretability of classification. Biometrics 1965, 21, 768–769.

[6] Al-Sultan, K. S. and Khan, M. M. Computational experience on four algorithms for the hard clustering problem. Pattern Recogn. Lett.1996, 17(3), 295–308.

[7] Hartigan, J.A. Clustering Algorithms, John Wiley and Sons,724 Inc., New York, NY, 1975.

[8] Theodoridis, S., Koutroumbas, K.: Pattern Recognition, 2nd edn. Elsevier Academic Press, Amsterdam 2003.

[9] Hruschka, E.R.; Campello, R.J.G.B.; Freitas, A.A. & de Carvalho, A.C.P.L.F. A Survey of Evolutionary Algorithms for Clustering. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews. 2009, 39(2), pp.133-155, March 2009, ISSN 1094-6977

[10] MacQueen, J.: Som methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkely Symposium on Mathematical Statistics and Probability, 1967, pp. 281–297.

[11] Selim, S.Z., Ismail, M.A. K-means type algorithms: a generalized convergence theorem and characterization of local optimality, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6 (1984) 81–87.

[12] Ball, G., Hall, D. A clustering technique for summarizing multivariate data. Behavioral Science 12, 1967, 153–155..

[13] Huang, K. A synergistic automatic clustering technique (Syneract) for multispectral image analysis. Photogrammetric Engineering and Remote Sensing 2002, 1(1), 33–40

[14] Rosenberger, C., Chehdi, K. Unsupervised clustering method with optimal estimation of the number of clusters: Application to image segmentation. In: Proc. IEEE International Conference on Pattern Recognition (ICPR), vol. 1, Barcelona, 2000, pp. 1656– 1659.

[15] Pelleg, D., Moore, A. X-means Extending K-means with efficient estimation of the number of clusters. In: Proceedings of the 17th International Conference on Machine Learning, 2000, pp. 727–734.

[16] Hamerly, G Learning structure and concepts in data using data clustering, PhD Thesis, University of California, San Diego, 2003.

[17] Halkidi, M., Batistakis, Y., Vazirgiannis, M. On clustering validation techniques. *Journal of Intelligent Information Systems* (JIIS) 2001, 17(2-3), 107–145.

[18] Bandyopadhyay, S., Maulik, U. Genetic clustering for automatic evolution of clusters and application to image classification. *Pattern Recognition* 2002, 35, 1197–1208.

[19] Venkatesh K., Satapathy, S. C, JVR Murthy, PVGD Prasad Reddy Hybridized Improved Genetic Algorithm with Variable Length Chromosome for Image Clustering. In *IJCSNS* , 2007, 7(11), 21-131.

[20] Das, S. and Konar, A. Automatic image pixel clustering with an improved differential evolution *Applied Soft Computing* 2009, (1) p. 226-236

[21] Bandyopadhyay, S. Simulated Annealing for Fuzzy Clustering: Variable Representation, Evolution of the Number of Clusters and remote Sensing Applications, 2003 unpublished, private communication

[22] Tseng, L. Y. and Yang, S. B. A genetic approach to the automatic clustering algorithm, *Pattern Recognition,* 2001, 34(2), pp. 415-424.

[23] Lin H. J., Yang, F. W. and Kao, Y. T. An efficient GA-based clustering technique, *Tamkang Journal of Science and Engineering,* 2005, 8(2), pp. 113-122.

[24] Lai, C. C., A novel clustering approach using hierarchical genetic algorithms, *Intelligent Automation and Soft Computing,* 2005, 11(3), pp. 143-153.

[25] Alexander V. Lukashin and Rainer Fuchs, Analysis of temporal gene expression profiles: Clustering by simulated annealing and determining the optimal number of clusters, *Bioinformatics*, 2000, 17(5), pp. 405–414.

[26] Das, S, Abraham, A and Konar, A Automatic Kernel Clustering with Multi-Elitist Particle Swarm Optimization Algorithm, *Pattern Recognition Letters*, Elsevier Science, 2008, 29, pp. 688-699.

[27] Liu, Wemping., Chih. Cheng. Hung, Bor. Chen. Kuo, and Tommy. Coleman, An Adaptive Clustering Algorithms Based on the Possibility Clustering and ISODATA for Multispectral Image Classification, proceedings of the International Society for Photogrammetry and Remote Sensing (ISPRS – XXI Congress), Beijing, China, 2008, July 3 – 11.

[28] Othman, R.M., Deris, S., Illias, R.M., Zakaria, Z. and Mohamad, S.M. Automatic clustering of gene ontology by genetic algorithm, Int'l J. Information Technology, 2006 3(1) pp. 37-46.

[29] Markus, M. and Grani, M. Automatic Cluster Number Selection Using a Split and Merge K-Means Approach 2009 20th International Workshop on Database and Expert Systems Application p. 363-367

[30] Das, S, Chowdhury,A, and Abraham, A. A Bacterial Evolutionary Algorithm for automatic data clustering Evolutionary Computation 2009 CEC 09 IEEE Congress on p. 2403-2410.

[31] Omran, M.G., Engelbrecht, A.P., Salman, A. Dynamic clustering using particle swarm optimization with application in image classification. In *Pattern Analysis and Application*. 2006, 332-344.

[32] Abraham, A., Das, S., and Roy, S. Swarm intelligence algorithms for data clustering in Soft Computing for Knowledge Discovery and Data Mining Book. New York: Springer-Verlag, , 2007, 279–313.

[33] Alam, S., Dobbie, G., Riddle, P. An Evolutionary Particle Swarm Optimization algorithm for data clustering. In Swarm Intelligence Symposium, 2008, 1-6.

[34] Kennedy, J., Eberhart, R.C. Particle swarm optimization. In: Proc. of the IEEE Int. Conf. on Neural Networks, Piscataway, NJ, 1995 1942–1948.

[35] Shi, Y., and Eberhart, R. C, A modified particle swarm optimizer. In Prc. IEEE Congr. Evol. Comput., 1998, 69–73.

[36] Turi, R. Clustering-based colour image segmentation, Ph.D. Thesis, Monash University, Australia, 2001.

[37] Srinivas, M. and Patnaik, L.M. Adaptive probabilities of crossover and mutation in genetic algorithms, *IEEE Transactions on Systems, Man and Cybernetics* 1994, 24 (4) , 656–667.

Low Power NoC Switch using Novel Adaptive Virtual Channels

Rabab Ezz-Eldin[1], Magdy A. El-Moursy[2] and Amr M. Refaat[3]

**[1]Electrical and Computer Engineering Department, Bani-suef University,
Bani-suef, Egypt, Electronics Research Institute,
Cairo, Egypt**

[2]Mentor Graphics Corporation, Cairo, Egypt

**[3]Electrical Engineering Department, Fayoum University,
Fayoum, Egypt**

Abstract

Adaptive Virtual Channel (AVC) is proposed as a novel technique to achieve low power NoC switch. Power supply gating is employed to reduce the power dissipation of NoC switch without degrading network performance. Hierarchical multiplexing tree is used to achieve efficient AVC. AVC could reduce both dynamic and leakage power of the switch. Hierarchical multiplexing tree decreases the area of the switch which reduces the dynamic power by 60%. Using the leakage power reduction technique, the average leakage power consumption of Adaptive Virtual Channels is reduced by up to 97%.

Keywords: *Virtual Channels, NoC, power gating, hierarchical multiplexing.*

1. Introduction

As the technology continuously scales down, the need for high performance, low power, as well as high throughput and reliable integrated circuits increases. Integrated Circuits are moving towards System on a Chip (SoC) which increases the circuit complexity. In recent years the complexity of interconnection architectures of the SoCs increased significantly. Network on Chip (NoC) was proposed as a solution for the interconnection problem. NoC is an on-chip network composed of processing cores connected by switches and communication channels [1]. Each physical channel can be split into several virtual channels using multiple parallel buffers. All virtual channels share the bandwidth of the physical channel. Different number of virtual channels is previously used to improve the network throughput. As the number of virtual channels increases, network throughput increases [2]. A tradeoff between power dissipation of the circuit and network throughput exits. In the previous network implementations, fixed number of virtual channels was used [3-5]. Achieving high throughput while reducing power dissipation is the objective of this paper.

Power consumption grows rapidly in NoC as interconnection complexity increases [6]. Reducing the power consumption becomes the first objective in NoC design. Power consumption should be minimized for reliability and cost-efficiency. Dynamic power and leakage power are the main components of power dissipation in NoC. Reducing leakage power is taking a lot of attention since it is dominating the power dissipation in today's and tomorrow's technologies. The main focus of this paper is to present a new Adaptive Virtual Channel (AVC) technique as a novel technique to reduce power dissipation of NoC switch. AVC allows efficient power gating to be employed to reduce power dissipation of the switch as shown in Figure 1. AVC is used to reduce the leakage power of a network switch. Hierarchical multiplexing tree is shown to be efficient in reducing not only the leakage power but also the dynamic power of the switch.

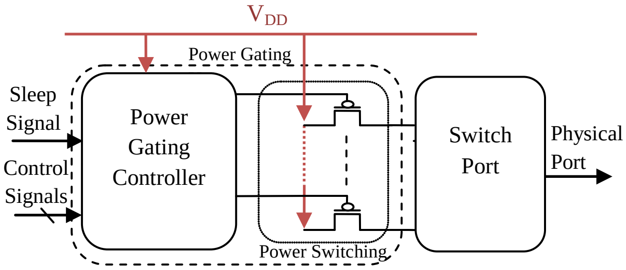


Figure 1: Block diagram of switch port with power gating

The paper is organized as follows: in section 2, AVC architecture is proposed. The power gating mechanism is presented in section 3. In section 4, simulation results are demonstrated. Conclusions are provided in section 5.

2. Adaptive Virtual Channel Architecture

Adaptive number of virtual channels is achieved in the proposed technique to multiplex the input channels to network switch. The characteristics of the network traffic are used as indicator to enable/disable the appropriate number of virtual channels. The number of available virtual channels is divided into power-of-two sets of configurable virtual channels. Each set could be configured as active or in-active. AVC multiplexing tree where the virtual channels are located at the leaves of the tree and the physical port is located at the root ($level_n$) is illustrated Figure 2. The tree is developed as a binary tree to optimize circuit implementation. Non-binary tree would complicate circuit implementation with limited flexibility of activating arbitrary number of virtual channels. The number of virtual channels equals 2^p where $p = m + n$. The number of connected virtual channels to a single cell (set of VC) is 2^m. n is the number of multiplexing levels. The maximum number of active sets in the tree is 2^n. m and n are positive integer numbers where $m \geq 1$ and $n \geq 0$. Each set is connected to one multiplexing cell located at the first multiplexing level. Every two cells in a low level of the tree are connected to one cell in the upper level. The total number of cells in the tree is given by

$$k = 2(2^n - 1) \qquad (1)$$

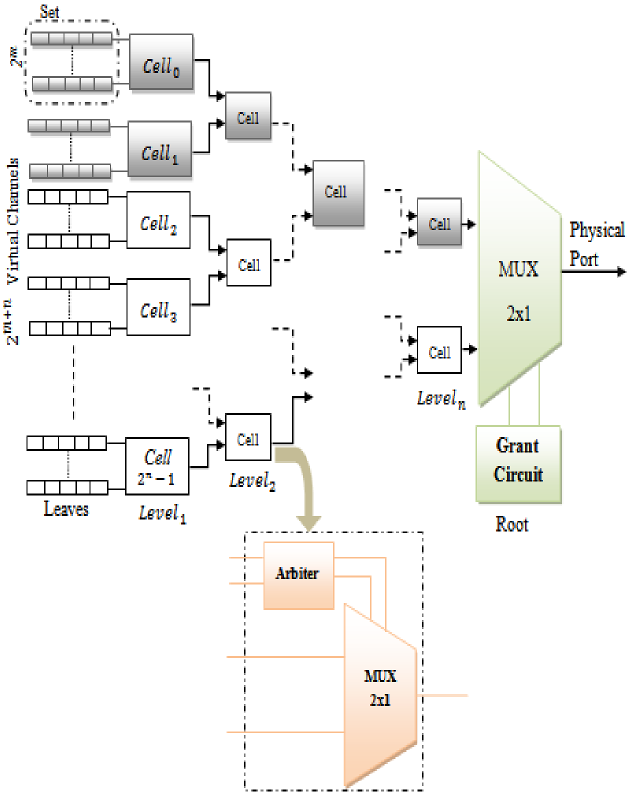


Figure 2: Adaptive Virtual Channel multiplexing tree structure

The root consists of one multiplexer 2x1 and one grant circuit 2x2 as shown in Figure 2. Every cell (in all levels expect $level_1$) of the tree consists of one arbiter 2x2 and one multiplexer 2x1. Cells in $level_1$ contains one multiplexer 2^mx1 and one arbiter 2^mx2^m. At the root, only one virtual channel is granted the physical port. m and n introduce a single degree of freedom in designing the switch. The tree structure can be created with p different implementation options. m and n defines a tradeoff between circuit delay and configurability. For n equals zero, the tree contains only the root. Therefore, no multiplexing tree is required. All virtual channels operate simultaneously. Eliminating the multiplexing hierarchy reduces the circuit delay. However, all virtual channels are included in the root. The flexibility of configuring the virtual channels is minimized and no saving in power is possible.

The delay of the tree structure increases with increasing the number of multiplexing levels n. However, the area of the switching circuitry decreases as n increases since the hardware implementation is optimized. The flexibility of activating/deactivating the virtual channel sets increases as n increases which allow saving in power components.

The virtual channels are activated in groups, according to the network traffic. Multiplexing tree activation is highlighted in Figure 2 for m =1 and for light traffic (*i.e.* only two sets are activated). All upstream cells connected to the active virtual channel sets are, accordingly, activated. In-active virtual channels are power gated to reduce the leakage power dissipation as described in section 3. n should be maximized to maximize the circuit configurability, minimize circuit area, and maximize the power saving as describe in section 4.

3. POWER GATING MECHANISM

In order to reduce power dissipation, virtual channels are deactivated when the network traffic is light. The hierarchical multiplexing structure is exploited to configure the virtual channel sets to active/in-active mode according to network traffic. Power supply gating is employed to deactivate the cells and the virtual channels connected to them. Each cell has one power gating switch. Deactivating the virtual channel sets reduces the leakage power dissipation of the whole switch without degrading the network throughput since it is applied according to traffic characteristics. Power management is performed using a power gating controller in addition to the sleep transistors. In section 3.1, the sleep controller unit is described. A mechanism to size the sleep transistor of a multiplexing cell is presented in section 3.2.

3.1 The Sleep Controller

A power gating control unit is used to control activating the virtual channels. The controller manages the sleep transistor of each cell. The sleep controller unit has two inputs and one output signal. A *cntrl_sleep* signal is used to enable/disable the switch port. This signal turns off the whole switch. The *nt* signal (*n* bits) indicates the status of the network traffic.

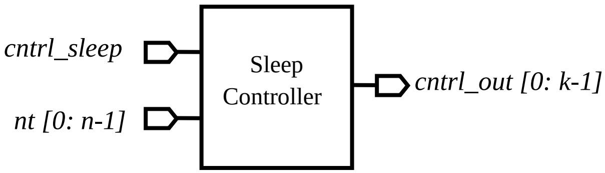


Figure 3: The sleep controller block diagram

It is required to find a general canonical representation of the truth table for the first level of the *cntrl_out* of the sleep controller. Taking into consideration that the number of bits for inputs and outputs change depending on the traffic heaviness and the number of controlled cells, cells are indexed as shown in Figure 2. Therefore, the upper cell in the $level_1$ of the tree has the lowest index of zero. The index increases going from top to bottom. The index of the bottom cell in $level_1$ is $(2^n - 1)$ as shown in Figure 2. The equation is algebraically expressed in a sum of minterms form

$$cntrl\_out_x(nt_0, nt_1, .. nt_{n-1}) = \sum(0,1, \dots \dots, x-1)$$

$$\text{where} \quad 1 \leq x \leq 2^n - 1, \qquad (2)$$

where x is the index of the cell. This equation produces $(2^n - 1)$ columns of the output truth table, each column controls one cell in the $level_1$. At $x = 0$, the column of $cntrl\_out_0$ equals zeroes which means always activate $Cell_0$ regardless the traffic heaviness. Turning on a child cell in $level_1$ requires turning on all parents cells of this child. Therefore, the truth table of every cell in $level_1$ which has even index is the same truth table of all parent cells in the same path from $level_1$ to the root.

The output signal *cntrl_out* has *k* bits depending on the number of cells in the switch port as shown in Figure 3. The *cntrl_out* signal is used to manage sleep transistors according to the required number of virtual channels to be activated. Depending on the value of the *cntrl_out* signal, some sleep transistors are switched ON to activate its connected cells and the other sleep transistors are switched OFF to ensure that the connected cells are deactivated.

When *cntrl_sleep* signal is 1, all the bits of the *cntrl_out* are 1 and hence all the sleep transistors will be switched OFF. Thereby the switch port is forced to turn off regardless of the value of the *nt*. This increases the power saving of the port as shown in section 4. When *cntrl_sleep* signal is 0, the sleep controller calculates the value of the *cntrl_out* signal according to input signal *nt* which activates certain number of virtual channels. Activating virtual channels depends on the traffic heaviness which can take different levels. The number of traffic heaviness levels depends on granularity of activating the virtual channels which equals 2^n. For example, for number of virtual channels of eight and for *m=1*, there are four levels of traffic heaviness (*n=2*), “Very Heavy”, ”Heavy”, ”Light”, “Very Light”. For *m=2*, there are only two levels of traffic heaviness, “Heavy” and ”Light”. With very heavy traffic profile, all virtual channels are activated by switching ON all cells.

The granularity of activating the virtual channels is 2^m. For *m=1*, binary multiplexing tree is used and two virtual channels are activated at a time. For *m=2*, four virtual channels are activated simultaneously. As *m* increases, the depth of the multiplexing tree decreases reducing the area overhead and the critical path delay. On the other hand, for small *m*, larger power saving is achieved since power gating could be applied with higher granularity. Hierarchical switching increases the flexibility of activating the virtual channels making the switch more adaptive to the changes in the traffic characteristics. Accordingly, reducing *m* increases the power saving.

3.2 Circuit Implementation of Power Switching

The power switching block consists of *k* sleep transistors. In our architecture, the sleep transistors are implemented by PMOS transistors to gate the power supply path to ground. Sleep transistor acts as a switch to turn-off the supply voltage during the sleep mode. On the other hand, sleep transistors in the active mode are ON and hence the value of the virtual supply node is v_{DD}. Sizing the sleep transistor affects both circuit performance as well as the efficiency of power saving.

A tradeoff in sizing the sleep transistor exists. During the active mode, the sleep transistor impedes the flow of the supply current requiring the transistor to be up sized to keep circuit performance. On the other hand, sizing up the sleep transistor reduces its ability to mitigate the leakage current and power. In addition, the power gating control circuit dissipates more dynamic power with larger sleep transistor.

The traffic heaviness signal *nt* is assumed to arrive to the target switch one clock cycle before the actual cycle at which the signal is needed to activate the cells. This assumption allows only one clock cycle to switch the cell from sleep-to-active mode. The switching time from sleep-to-active TSA must be less than or equal to the critical path delay t_d of the cell circuitry.

$$TSA \leq t_d \tag{3}$$

The cell is considered active when its virtual v_{DD} node reaches 90% of v_{DD}. Sizing sleep transistor and its implication on the circuit performance is demonstrated in section 4.

4. SIMULATION RESULTS

The proposed architecture is implemented using the ADS tools. 45nm technology is used with supply voltage of 1V. A switch port with eight virtual channels is considered. The tradeoff between the TSA switching time, the critical path delay of the cell, and the leakage power reduction is presented in Figure 4. Leakage power increases with sizing up the sleep transistor. To increase power saving, the sleep transistor needs to be sized down. On the other hand, TSA could not be larger than t_d since only one clock cycle is needed to switch the cell from sleep-to-active mode. The intersection point on t_d and TSA curves is used as the optimum size for high performance and low power switch design. Based on that, the width of sleep transistor is determined to be 0.35µm for the target technology.

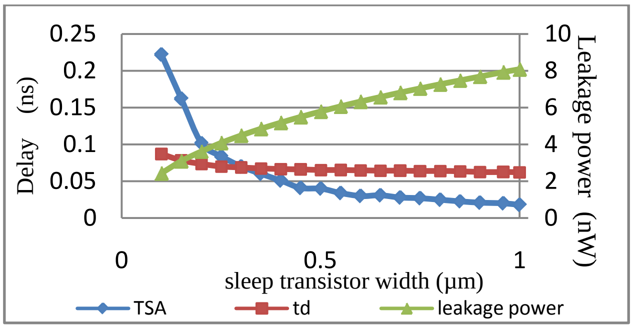


Figure 4: Sleep mode leakage power, critical path delay and TSA for different sleep transistor widths

4.1 Depth of the multiplexing tree

For eight virtual channels, there are p =3 implementation options

$$\begin{cases} option\ A: & m=3, & n=0, & k=0 \\ option\ B: & m=2, & n=1, & k=2 \\ option\ C: & m=1, & n=2, & k=6 \end{cases} \tag{4}$$

For option A, The tree structure consists of only the root. For option B, the available virtual channels are divided into two sets using four virtual channels per set. For option C, the available virtual channels are divided into four sets using binary multiplexing tree and two multiplexing levels. There are a total of six cells in the tree where at least two can be, simultaneously, activated.

The required area to implement the three options, including the area of the sleep controller and sleep transistors, is shown in Figure 5. The area of multiplexing tree of option *C* is less than the area of the multiplexing tree of option *A* and *B*. As compared to option *A*, the area decreases in option *B* by 50.93% and by 60.11% for option *C*. The overhead in the input gate capacitance (sleep controller and sleep transistors) in option *C* is 6.6% of the total port capacitance. In option *B*, the overhead is only 3.61%.

The hierarchical tree implementation has two-fold effect in reducing power dissipation of the switch. The dynamic power is reduced for the reduction in the input gate capacitance of the switch. With hierarchical multiplexing, dynamic power could decrease by up to 60%. In addition, the leakage power is decreased with light traffic since power gating is more efficient.

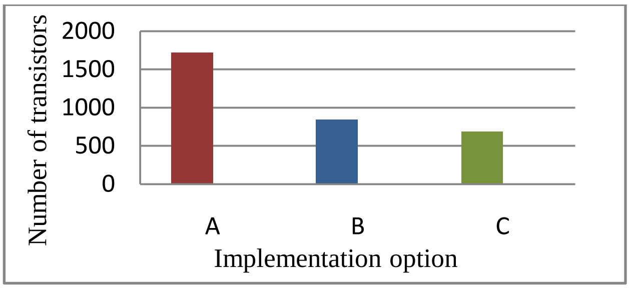


Figure 5: The area of switch port for different number of virtual channel per one set

On the other hand, the hierarchical tree structure increases the critical path delay of the circuit reducing the maximum operation frequency. The maximum operation frequency and leakage power for the three implementation options are listed in Table 1. The maximum operation frequency and leakage power decrease with increasing the number of levels. The leakage power for option *C* reduces by 87.12 % as compare to the leakage power of option *A*. A pipeline stage could be used to maintain the operating frequency but latency of switching would increase.

Table 1. Maximum operating frequency and leakage power for three implementation options with different hierarchical depths

m	*Maximum operation frequency*		*Maximum leakage power*	
	(GHz)	*Reduction (%)*	*(nW)*	*Reduction (%)*
3	18.99	-	2821.11	-
2	12.18	35.86	616.85	78.13
1	8.44	55.76	363.32	87.12

4.2 Power dissipation of AVC

For total number of virtual channels of eight (*p* =3), *m =1* and *n*=2. There are four levels of traffic heaviness listed in Table 2.The network traffic is used to control the number of active virtual channels. Since *n* =2, two, four, six or eight virtual channels could be simultaneously activated depending on the traffic of the network.

The reduction in leakage power dissipation is reported in Table 2 for different network traffic characteristics. The power saving increases as the number of active virtual channels decreases. Power saving could reach up to 81% when only two virtual channels are simultaneously activated. When no virtual channels are activated, power saving increases up to 97%. Adaptive virtual channel with hierarchical multiplexing tree significantly decreases the power consumption of the switch.

Table 2. the leakage power saving with different number of virtual channels

Number of Virtual channel	***6***	***4***	***2***	***0***
Traffic heaviness	Heavy	Light	Very Light	No traffic
Power saving (%)	17.3	34.1	80.2	96.8

5. CONCLUSIONS

Adaptive Virtual Channel is proposed as an efficient novel technique to reduce power dissipation of NoC switch. AVC uses hierarchical multiplexing tree and power gating mechanism to reduce both dynamic and leakage power dissipation of the switch. The virtual channels are activated based on the network traffic. The area of switch port reduces with increasing hierarchical levels which decreases the dynamic power by up to 60%. Power saving increases with decreases the number of active virtual channels. The reduction in leakage power dissipation could reach 81% when only two virtual channels are activated simultaneously using AVC. At in-active mode of the switch port, power saving could increase up to 97%.

References

[1] Fernando Moraes, Ney Calazans, Aline Mello, Leandro Möller, Luciano Ost,"HERMES: an Infrastructure for Low Area Overhead Packet-switching Networks on Chip", Integration, the VLSI Journal, Vol.38 (1), October 2004, pp. 69-93.

[2] William J. Dally, "Virtual-Channel Flow Control", IEEE Transactions on parallel and distributed systems, Vol. 3, No.2, March 1992, pp. 194-205.

[3] Partha Pratim Pande, Cristian Grecu, Michael Jones,Andre ´ Ivanov, Resve Saleh, "Performance Evaluation and Design Trade-Offs for Network-on-Chip Interconnect Architectures", IEEE Transactions on computers, Augusts 2005, pp. 1025-1040.

[4] Partha Pratim Pande, Cristian Grecu, André Ivanov, Res Saleh, "High-Throughput Switch-Based Interconnect for Future SoCs", In Proceedings of the 3rd IEEE International workshop on SoC for real-time applications, July. 2003, pp. 304-310.

[5] Aline Mello, Leonel Tedesco, Ney Calazans, Fernando Moraes, "Virtual Channels in Networks on Chip: Implementation and Evaluation on Hermes NoC", In Proceedings of the Integrated Circuits and Systems Design, September 2005, pp. 178 – 183.

[6] Srinivasan Murali, David Atienza, Paolo Meloni, Salvatore Carta, Luca Benini, Giovanni De Micheli, Luigi Raffo, "Synthesis of Predictable Networks-on-Chip-Based Interconnect Architectures for Chip Multiprocessors.", IEEE Transactions on Very Large Scale Integration (VLSI) Systems, Vol. 15, No. 8, August 2007, pp. 869-880.

Rabab Ezz-Eldien received the B.Sc. degree in Electronics and communications department with honors from the Faculty of Engineering, Fayoum University, Fayoum, Egypt, in 2004. She is currently working a research assistant in Electrical and Computer department at Bani-suef University. She is joined the M.Sc program in Fayoum university in 2009. Her areas of interest include Networks-on-Chip, Computer Architecture and Embedded System.

Magdy A. El-Moursy was born in Cairo, Egypt in 1974. He received the B.S. degree in electronics and communications engineering (with honors) and the Master's degree in computer networks from Cairo University, Cairo, Egypt, in 1996 and 2000, respectively, and the Master's and the Ph.D. degrees in electrical engineering in the area of high-performance VLSI/IC design from University of Rochester, Rochester, NY, USA, in 2002 and 2004, respectively. In summer of 2003, he was with STMicroelectronics, Advanced System Technology, San Diego, CA, USA. Between September 2004 and September 2006 he was a Senior Design Engineer at Portland Technology Development, Intel Corporation, Hillsboro, OR, USA. During September 2006 and February 2008 he was assistant professor in the Information Engineering and Technology Department of the German University in Cairo (GUC), Cairo, Egypt. Dr. El-Moursy is currently a Technical Lead in the Mentor Hardware Emulation Division, Mentor Graphics Corporation, Cairo, Egypt. His research interest is in Networks-on-Chip, interconnect design and related circuit level issues in high performance VLSI circuits, clock distribution network design, and low power design. He is the author of more than 30 papers, four book chapters, and one book in the fields of high speed and low power CMOS design techniques and high speed interconnect.

Amr M. Gody; Joined Cairo University, faculty of Engineering in 1986. He is earned BSc. in Electronics and communication engineering in 1991 with an honor degree. He is earned the M.Sc degree in Electronics and communication engineering in 1995 from Cairo University, faculty of Engineering. He is joined the PhD program in Cairo university in 1996. He is earned the PhD in 1999 in the field of speech signal processing. Amr is Associate professor in Fayoum University, Electrical engineering department and he is acting as head of Electrical Engineering since 2010 till now.

Image Mining for Mammogram Classification by Association Rule Using Statistical and GLCM features

Aswini kumar mohanty[1], Sukanta kumar swain[2] ,Pratap kumar champati[3] ,Saroj kumar lenka[4]

[1] Phd scholar, SOA University , Bhubaneswar,
Orissa, India

[2]NIIS,Madanpur,Bhubaneswar,
Orissa,India

[3] Deptt.Comp.Sc,.ABIT,Cuttack,
Orissa,India

[4] Mody Univesity,Department of Comp Sc,Laxmangargh,
rajstan,India

Abstract

The image mining technique deals with the extraction of implicit knowledge and image with data relationship or other patterns not explicitly stored in the images. It is an extension of data mining to image domain. The main objective of this paper is to apply image mining in the domain such as breast mammograms to classify and detect the cancerous tissue. Mammogram image can be classified into normal, benign and malignant class and to explore the feasibility of data mining approach. A new association rule algorithm is proposed in this paper. Experimental results show that this new method can quickly discover frequent item sets and effectively mine potential association rules. A total of 26 features including histogram intensity features and GLCM features are extracted from mammogram images. A new approach of feature selection is proposed which approximately reduces 60% of the features and association rule using image content is used for classification. The most interesting one is that oscillating search algorithm which is used for feature selection provides the best optimal features and no where it is applied or used for GLCM feature selection from mammogram. Experiments have been taken for a data set of 300 images taken from MIAS of different types with the aim of improving the accuracy by generating minimum no. of rules to cover more patterns. The accuracy obtained by this method is approximately 97.7% which is highly encouraging.

Keywords: *Mammogram; Gray Level Co-occurrence Matrix feature; Histogram Intensity; Genetic Algorithm; Branch and Bound technique; Association rule mining*.

1. Introduction

Breast Cancer is one of the most common cancers, leading to cause of death among women, especially in developed countries. There is no primary prevention since cause is still not understood. So, early detection of the stage of cancer allows treatment which could lead to high survival rate. Mammography is currently the most effective imaging modality for breast cancer screening. However, 10-30% of breast cancers are missed at mammography [1]. Mining information and knowledge from large database has been recognized by many researchers as a key research topic in database system and machine learning Researches that use data mining approach in image learning can be found in [2,3].

Data mining of medical images is used to collect effective models, relations, rules, abnormalities and patterns from large volume of data. This procedure can accelerate the diagnosis process and decision-making. Different methods of data mining have been used to detect and classify anomalies in mammogram images such as wavelets [4,5], statistical methods and most of them used feature extracted using image processing techniques [6].Some other methods are based on fuzzy theory [7,8] and neural networks [9]. In this paper we have used classification method called Decision tree classifier for image classification [10-12].

Classification process typically involves two phases: training phase and testing phase. In training phase the properties of typical image features are isolated and based on this training class is created .In the subsequent testing phase , these feature space partitions are used to classify the image. A block diagram of the method is shown in figure1.

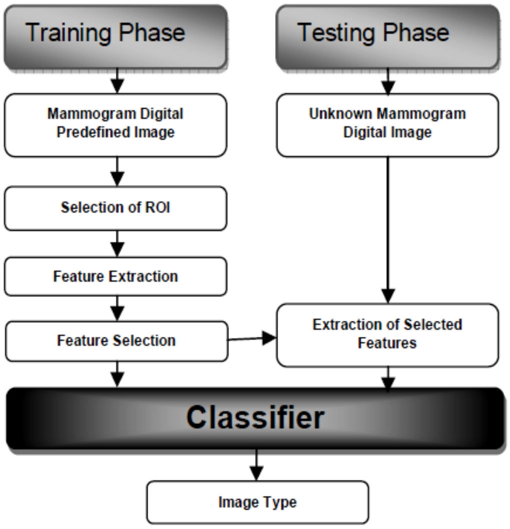


Fig.1.Block diagram for mammogram classification system

We have used association rule mining using image content method by extracting low level image features for classification. The merits of this method are effective feature extraction, selection and efficient classification. The rest of the paper is organized as follows. Section 2 presents the preprocessing and section 3 presents the feature extraction phase. Section 4 discusses the proposed method of Feature selection and classification. In section5 the results are discussed and conclusion is presented in section 6.

2. Methodologies

2.1 Digital mammogram database

The mammogram images used in this experiment are taken from the mini mammography database of MIAS (http://peipa.essex.ac.uk/ipa/pix/mias/). In this database, the original MIAS database are digitized at 50 micron pixel edge and has been reduced to 200 micron pixel edge and clipped or padded so that every image is 1024 X 1024 pixels. All images are held as 8-bit gray level scale images with 256 different gray levels (0-255) and physically in portable gray map (pgm) format. This study solely concerns the detection of masses in mammograms and, therefore, a total of 100 mammograms comprising normal, malignant and benign case were considered. Ground truth of location and size of masses is available inside the database.

2.2. Pre-processing

The mammogram image for this study is taken from Mammography Image Analysis Society (MIAS)†, which is an UK research group organization related to the Breast cancer investigation [13]. As mammograms are difficult to interpret, preprocessing is necessary to improve the quality of image and make the feature extraction phase as an easier and reliable one. The calcification cluster/tumor is surrounded by breast tissue that masks the calcifications preventing accurate detection and shown in Figure.3. .A pre-processing; usually noise-reducing step [14] is applied to improve image and calcification contrast figure 3. In this work the efficient filter (CLAHE) was applied to the image that maintained calcifications while suppressing unimportant image features. Figures 3 shows representative output image of the filter for a image cluster in figure 2. By comparing the two images, we observe background mammography structures are removed while calcifications are preserved. This simplifies the further tumor detection step.

.Contrast limited adaptive histogram equalization (CLAHE) method seeks to reduce the noise produced in homogeneous areas and was originally developed for medical imaging [15]. This method has been used for enhancement to remove the noise in the pre-processing of digital mammogram [16]. CLAHE operates on small regions in the image called tiles rather than the entire image. Each tile's contrast is enhanced, so that the histogram of the output region approximately matches the uniform distribution or Rayleigh distribution or exponential distribution. Distribution is the desired histogram shape for the image tiles. The neighbouring tiles are then combined using bilinear interpolation to eliminate artificially induced boundaries. The contrast, especially in homogeneous areas, can be limited to avoid amplifying any noise that might be present in the image.The block diagram of pre-processing is shown in Figure 4.

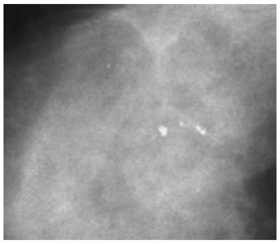


Fig. 2 ROI of a Benign

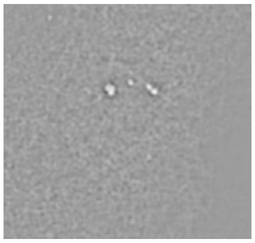


Fig. 3 ROI after Pre-processing Operation

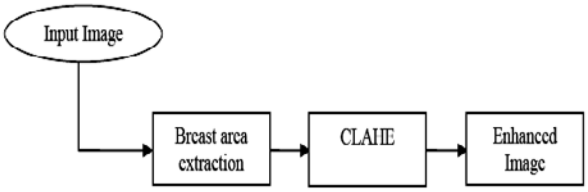


Fig.4. Image pre-processing block diagram.

2.3. Histogram Equalization

Histogram equalization is a method in image processing of contrast adjustment using the image's histogram [17]. Through this adjustment, the intensities can be better distributed on the histogram. This allows for areas of lower local contrast to get better contrast. Histogram equalization accomplishes this by efficiently spreading out the most frequent intensity values. The method is useful in images with backgrounds and foregrounds that are both bright or both dark. In particular, the method can lead to better views of bone structure in x-ray images, and to better detail in photographs that are over or under-exposed. In mammogram images Histogram equalization is used to make contrast adjustment so that the image abnormalities will be better visible.

3. Feature extraction

Features, characteristics of the objects of interest, if selected carefully are representative of the maximum relevant information that the image has to offer for a complete characterization a lesion [18, 19]. Feature extraction methodologies analyze objects and images to extract the most prominent features that are representative of the various classes of objects. Features are used as inputs to classifiers that assign them to the class that they represent.

In this Work intensity histogram features and Gray Level Co-Occurrence Matrix (GLCM) features are extracted.

3.1 Intensity Histogram Features

Intensity Histogram analysis has been extensively researched in the initial stages of development of this algorithm [18,20]. Prior studies have yielded the intensity histogram features like mean, variance, entropy etc. These are summarized in Table 1 Mean values characterize individual calcifications; Standard Deviations (SD) characterize the cluster. Table 2 summarizes the values for those features.

Table 1: Intensity histogram features

Feature Number assigned	***Feature***
1.	Mean
2.	Variance
3.	Skewness
4.	Kurtosis
5.	Entropy
6.	Energy

In this paper, the value obtained from our work for different type of image is given as follows

Table 2: Intensity histogram features and their values

Image Type	***Features***					
	Mean	Variance	Skewness	Kurtosis	Entropy	Energy
normal	7.2534	1.6909	-1.4745	7.8097	0.2504	1.5152
malignant	6.8175	4.0981	-1.3672	4.7321	0.1904	1.5555
benign	5.6279	3.1830	-1.4769	4.9638	0.2682	1.5690

3.2 GLCM Features

It is a statistical method that considers the spatial relationship of pixels is the gray-level co-occurrence matrix (GLCM), also known as the gray-level spatial dependence matrix [21,22]. By default, the spatial relationship is defined as the pixel of interest and the pixel to its immediate right (horizontally adjacent), but you can specify other spatial relationships between the two pixels.

[†] peipa.essex.ac.uk/info/mias.html
.

Each element (*I*, *J*) in the resultant GLCM is simply the sum of the number of times that the pixel with value *I* occurred in the specified spatial relationship to a pixel with value *J* in the input image.

3.2.1 GLCM Construction

GLCM is a matrix **S** that contains the relative frequencies with two pixels: one with gray level value i and the other with gray level j-separated by distance d at a certain angle θ occurring in the image. Given an image window W(x, y, c), for each discrete values of d and θ, the GLCM matrix **S**(i, j, d, θ) is defined as follows.

An entry in the matrix **S** gives the number of times that gray level i is oriented with respect to gray level j such that $W(x_1, y_1)$=i and $W(x_2, y_2)$=j, then

$$(x_2, y_2) = (x_1, y_1) + (d\cos\theta,\ d\sin\theta)$$

We use two different distances d={1, 2} and three different angles θ={0°, 45°, 90°}. Here, angle representation is taken in clock wise direction.

Example

Intensity matrix

$$\begin{vmatrix} 1 & 3 & 1 & 1 & 1 \\ 2 & 2 & 4 & 2 & 1 \\ 1 & 4 & 1 & 4 & 1 \\ 2 & 2 & 2 & 1 & 1 \\ 1 & 1 & 2 & 2 & 1 \end{vmatrix} \text{for } \theta = 45^\circ \text{ and } d = 1$$

$$\begin{bmatrix} 3 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \end{bmatrix} \text{ for } \theta = 45^\circ \text{ and } d = 2 .$$

The Following GLCM features were extracted in our research work:

Autocorrelation, Contrast, Correlation, Cluster Prominence, Cluster Shade, Dissimilarity Energy, Entropy, Homogeneity, Maximum probability, Sum of squares, Sum average, Sum variance, Sum entropy, Difference variance, Difference entropy, information measure of correlation1, information measure of correlation2, Inverse difference normalized. Information difference normalized. The value obtained for the above features from our work for a typical image is given in the following table 3.

..

Table 3 : GLCM Features and values Extracted from Mammogram Image(Malignant)

Feature No	Feature Name	Feature Values
1	Autocorrelation	44.1530
2	Contrast	1.8927
3	Correlation	0.1592
4	Cluster Prominence	37.6933
5	Cluster Shade	4.2662
6	Dissimilarity	0.8877
7	Energy	0.1033
8	Entropy	2.6098
9	Homogeneity	0.6645
10	Maximum probability	0.6411
11	Sum of squares	0.1973,
12	Sum average	44.9329
13	Sum variance	13.2626
14	Sum entropy	133.5676
15	Difference variance	1.8188
16	Difference entropy	1.8927
17	Information measure of correlation1	1.2145
18	Information measure of correlation2	-0.0322
19	Inverse difference normalized	0.2863
20	Information difference normalized	0.9107

4. Feature subset selection

Feature subset selection helps to reduce the feature space which improves the prediction accuracy and minimizes the computation time [23]. This is achieved by removing irrelevant, redundant and noisy features .i.e., it selects the subset of features that can achieve the best performance in terms of accuracy and computation time. It performs the Dimensionality reduction.

Features are generally selected by search procedures. A number of search procedures have been proposed. Popularly used feature selection algorithms are Sequential forward Selection, Sequential Backward Selection, Genetic Algorithm and Particle Swarm Optimization, Branch and Bound feature optimization. In this work a new approach of oscillating search for feature selection technique [24] is proposed to select the optimal features. The selected optimal features are considered for classification. The oscillating search has been fully exploited to select the feature from mammogram which is one of the best techniques to optimize the features among many features. We have attempted to optimize the feature of GLSM and statistical features.

4.1 Oscillating Search Algorithms for Feature Selection

A new sub-optimal subset search method for feature selection is introduced. As opposed to other till now known subset selection methods the oscillating search is not dependent on pre-specified direction of search (forward or backward). The generality of oscillating search concept allowed us to define several different algorithms suitable for different purposes. We can specify the need to obtain good results in very short time, or let the algorithm search more thoroughly to obtain near-optimum results. In many cases the oscillating search over-performed all the other tested methods. The oscillating search may be restricted by a preset time-limit, what makes it usable in real-time systems.

Note that most of known suboptimal strategies are based on step-wise adding of features to initially empty features set, or step-wise removing features from the initial set of all features, Y. One of search directions, *forward* or *backward*, is usually preferred, depending on several factors [25], the expected difference between the original and the final required cardinality being the most important one. Regardless of the direction, it is apparent, that all these algorithms spend a lot of time testing features subsets having cardinalities far distant from the required cardinality *d*.

Before describing the principle of oscillating search, let us introduce the following notion: the "worst" features *o*-tuple in X_d should be ideally such a set$\bar{W} \subset X_d$, that
$J(X_d \setminus \bar{W}) = \max_{\mathcal{W} \subset X_d, |\mathcal{W}|=o} J(X_d \setminus \mathcal{W})$.

The "best" feature *o*-tuple for X_d should be ideally such a set $\bar{B} \subset Y \setminus X_d$, that
$J(X_d \cup B) = \max_{\mathcal{B} \subset Y \setminus X_d, |\mathcal{B}|=o} J(X_d \cup \mathcal{B})$.

In practice we allow also suboptimal finding of the "worst" and "best" *o*-tuples to save the computational time (see later).

4.1.1. Oscillating Search

Unlike other methods, the *oscillating search* (OS) is based on repeated modification of the current subset X_d of *d* features. This is achieved by alternating the *down*- and *up-swings*. The *down-swing* removes *o* "worst" features from the current set X_d to obtain a new set X_{d-o} at first, then adds *o* "best" ones to X_{d-o} to obtain a new current set X_d . The *up-swing* adds *o* "good" features to the current set X_d to obtain a new set X_{d+o} at first, then removes *o* "bad" ones from X_{d+o} to obtain a new current set X_d again. Let us denote two successive opposite swings as an *oscillation cycle*. Using this notion, the oscillating search consists of repeating oscillation cycles. The value of *o* will be denoted *oscillation cycle depth* and should be set to 1 initially. If the last oscillation cycle did not find better subset X_d of *d* features, the algorithm increases the oscillation cycle depth by letting *o* = *o*+1. Whenever any swing finds better subset X_d of *d* features, the depth value *o* is restored to 1. The algorithm stops, when the value of *o* exceeds the user-specified *limit* Δ. The course of oscillating search is illustrated on picture 1.

Every oscillation algorithm assumes the existence of some initial set of *d* features. Obtaining such an initial set will be denoted as an *initialization.* Oscillating algorithms may be initialized in different ways; the simplest ways are random selection or the SFS procedure. From this point of view the oscillating search may serve as a mechanism for tuning solution obtained in another way.

Whenever a feature *o*- tuple is to be added (or removed) from the current subset in the till now known methods, one of two approaches is usually utilized: the *generalized* adding (or moving) find s the optimum *o*-tuple by means of exhaustive search (e.g. in GSFS, GSBS) or the *successive adding (or removing) single features o-time* (e.g. in basic Plus-l Minus-r), which may fail to find the optimum *o*- tuple, but is significantly faster.

In fact, we may consider finding feature *o*-tuples to be equivalent to the feature selection problem, though at a "Second level". Therefore, we may use any search strategy for findings feature o-tuples. Because of proved effectiveness of floating search strategies we adopted the floating search principle as the third approach to adding (or removing) feature o-tuples in oscillating search. Note that in such a way one basic idea has resulted in defining a couple of new algorithms, as shown in the sequel.

For the sake of simplicity, let us denote the adding of feature o-tuple by ADD(o) and the removing of feature o-tuple by REMOVE(o). Finally, we introduce three versions of oscillating algorithm.

1. *Sequential oscillating search :* ADD (*o*) represents a sequention of *o* successive SFS steps (see [1]), REMOVE(*o*) represents a sequention of *o* successive SBS steps.
2. *Floating oscillating search :* ADD (*o*) represents adding of *o* features by means of the SFFS procedure (see [5]), REMOVE (*o*) represents removing of *o* features by means of the SFBS procedure.
3. *Generalized oscillating search :* ADD (*o*) represents adding of *o* features by means by means of the GSFS(*o*) represents removing of *o* features by means of the GSFS (*o*) procedure.

Remark : c serves as a swing counter.

Step 1 : *Initialization:* by means of the SFS procedure (or otherwise) find the initial set X_d of *d* features. Let *c* = 0, Let *o* = 1.

Step 2 : *Down-swing:* By means of the REMOVE (*o*) step remove the "worst" feature o-tuple from X_d to form a new set X_{d-o} (* if the J (X_{d-o}) value is not the so far best one among subsets having cardinality d-o, stop the down-swing and go to Step 3 *). By means of the ADD(o) step add the "best" feature o-tuple from Y\X_{d-o} to X_{d-o} to form a new subset $X^1{}_d$. If the J($X^1{}_d$) value is the so far best one among subsets having required cardinality d, let $X_d = X^1{}_d$, $c= 0$ and $o = 1$ and got **Step 4**.

Step 3 : *Last swing did not find better solution.* Let $c = c + 1$. If $c =2$, the none of previous two swings has found better solution; extend the search, by letting $o = o +1$. If $o > \Delta$, stop the algorithm, otherwise let c = 0.

Step 4 : Up-swing : By means of the ADD(o) step add the "best" feature o-tuple from $Y \setminus X_d$ *to* X_d to form a new set X_{d+0} (* If the $J\ (X_{d+o})$ value is not the so-far best one among subsets having cardinality $d+o$, stop th up-swing and go to **Step 5.** *). "Worst" feature o-tuple from X_{d+o} to from a new set $X^1{}_d$. If the J ($X^1{}_d$) value is the so far best one among subsets having required cardinality d, let $X_d = X^1{}_d$ $c=0$ and $o=1$ and go to **Step 2**.

Step 5 : Last swing did not find better solution : Let $c =c+1$. If $c =2$, the none of previous two swings has found better solution; extend the search by letting $o = o +1$, *If* $o > \Delta$, stop the algorithm, otherwise let c = 0 and go to **Step 2**

Remark : Parts of code enclosed in (* and *) brackets may be omitted to obtain a bit slower, more through procedure.

The algorithm is described also by a float chart on picture 2. The three introduced algorithm versions differ in their effectiveness and time complexity. The *generalized oscillating search* give the best results, but its use is restricted due to the time of complexity of generalized steps (especially for higher Δ). The *floating oscillating search* is suitable for findings solutions as close to optimum as possible in a reasonable time even in high-dimensional problems. The *sequential oscillating search* is the fastest but possibly the least effective algorithm versions with respect to approaching the optimal solution.

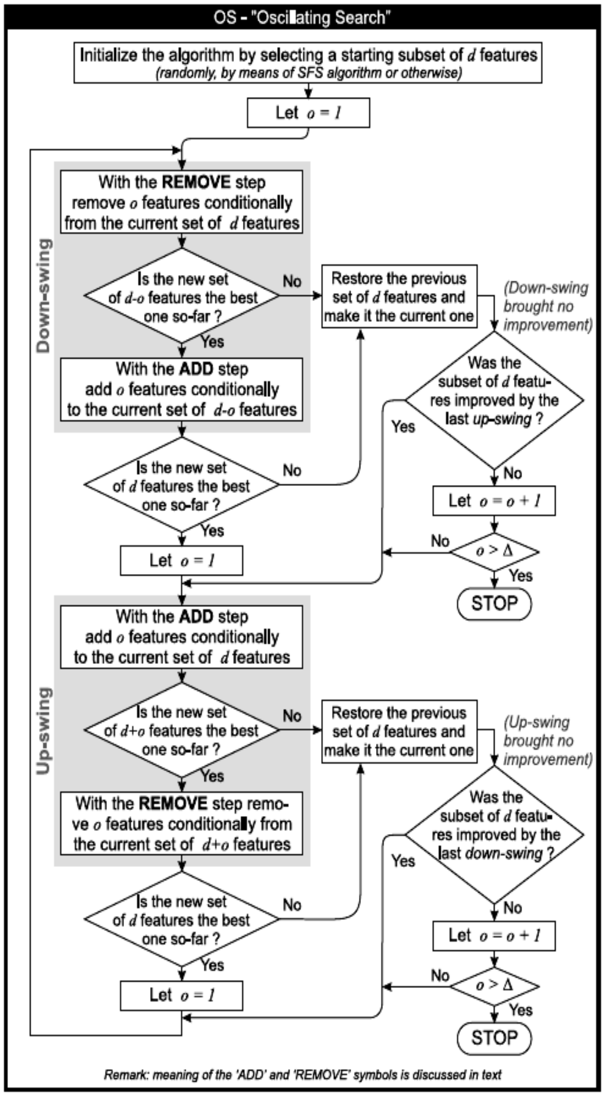

OS - "Oscillating Search"
Initialize the algorithm by selecting a starting subset of d features
(randomly, by means of SFS algorithm or otherwise)
Let o = 1
Down-swing
With the REMOVE step remove o features conditionally from the current set of d features
Is the new set of d-o features the best one so-far ?
No
Yes
Restore the previous set of d features and make it the current one
(Down-swing brought no improvement)
With the ADD step add o features conditionally to the current set of d-o features
Is the new set of d features the best one so-far ?
No
Yes
Was the subset of d features improved by the last up-swing ?
Yes
No
Let o = o + 1
No
o > Δ
Yes
STOP
Let o = 1
Up-swing
With the ADD step add o features conditionally to the current set of d features
Is the new set of d+o features the best one so-far ?
No
Yes
Restore the previous set of d features and make it the current one
(Up-swing brought no improvement)
With the REMOVE step remove o features conditionally from the current set of d+o features
Is the new set of d features the best one so-far ?
No
Yes
Was the subset of d features improved by the last down-swing ?
Yes
No
Let o = o + 1
No
o > Δ
Yes
STOP
Let o = 1
Remark: meaning of the 'ADD' and 'REMOVE' symbols is discussed in text

Fig.5. Simplified OS algorithm flowchart

By applying the proposed algorithm, it will produce a feature set contain best set of features which is less than the original set. This method will be providing a better and concrete relevant feature selection from 26 nos. of features to minimize the classification time and error and productive results in conjunction with better accuracy positively. The features selected by the method are given in table 4.

Table 4. Feature selected by proposed method

S.no.	Features
1	Cluster prominence
2	Energy
3	Information measure of correlation
4	Inverse difference Normalized
5	Skewness

6	Kurtosis
7	Contrast
8	Mean
9	Variance
10	Homogeneity
11	Entropy

5. Classification

5.1 Preparation of Transactional Database:

The selected features are organized in a database in the form of transactions [26], which in turn constitute the input for deriving association rules. The transactions are of the form[Image ID, *F*1*; F*2*; :::; F9*] where *F*1*:::F9*] are *9*features extracted for a given image.

5.2 Association Rule Mining:

Discovering frequent item sets is the key process in association rule mining.

In order to perform data mining association rule algorithm, numerical attributes should be discretized first, i.e. continuous attribute values should be divided into multiple segments. Traditional association rule algorithms adopt an iterative method to discovery, which requires very large calculations and a complicated transaction process. Because of this, a new association rule algorithm is proposed in this paper. This new algorithm adopts a Boolean vector method to discovering frequent item sets. In general, the new association rule algorithm consists of four phases as follows:

1. Transforming the transaction database into the Boolean matrix.
2. Generating the set of frequent 1-itemsets L1.
3. Pruning the Boolean matrix.
4. Generating the set of frequent k-item sets Lk(k>1).

The detailed algorithm, phase by phase, is presented below:

1. *Transforming the transaction database into the Boolean matrix:* The mined transaction database is *D*, with *D* having m transactions and *n* items. Let T={T1,T2,…,Tm} be the set of transactions and I={I1,I2,…,In}be the set of items. We set up a Boolean matrix Am*n, which has m rows and n columns. Scanning the transaction database *D*, we use a binning procedure to convert each real valued feature into a set of binary features. The 0 to 1 range for each feature is uniformly divided into k bins, and each of *k* binary features record whether the feature lies within corresponding range.

2. *Generating the set of frequent 1-itemset L1:* The Boolean matrix Am*n is scanned and support numbers of all items are computed. The support number Ij.supth of item Ij is the number of '1s' in the jth column of the Boolean matrix Am*n. If Ij.supth is smaller than the minimum support number, itemset {Ij} is not a frequent 1-itemset and the jth column of the Boolean matrix Am*n will be deleted from Am*n. Otherwise itemset {Ij} is the frequent 1-itemset and is added to the set of frequent 1-itemset L1. The sum of the element values of each row is recomputed, and the rows whose sum of element values is smaller than 2 are deleted from this matrix.

3. *Pruning the Boolean matrix:* Pruning the Boolean matrix means deleting some rows and columns from it. First, the column of the Boolean matrix is pruned according to Proposition 2. This is described in detail as: Let I• be the set of all items in the frequent set LK-1, where k>2. Compute all |LK-1(j)| where j belongs to I2, and delete the column of correspondence item j if |*LK* – 1(*j*)| is smaller than *k* – 1. Second, re-compute the sum of the element values in each row in the Boolean matrix. The rows of the Boolean matrix whose sum of element values is smaller than k are deleted from this matrix.

4. *Generating the set of frequent k-itemsets Lk:* Frequent k-item sets are discovered only by "and" relational calculus, which is carried out for the k-vectors combination. If the Boolean matrix *Ap*q* has q columns where 2 < *q* £ *n* and *minsupt*h £ *p* £ *m*, *k q c*, combinations of k-vectors will be produced. The 'and' relational calculus is for each combination of k-vectors. If the sum of element values in the "and" calculation result is not smaller than the minimum support number *minsupth*, the k-itemsets corresponding to this combination of kvectors are the frequent k-itemsets and are added to the set of frequent k-itemsets Lk.

6. Experimental results

In this paper we used association rule mining using image contents for the classification of mammograms. The average accuracy is 97.67 %. We have used the precision and recall measures as the evaluation metric for mammogram classification. Precision is the fraction of the number of true positive predictions divided by the total number of true positives in the set. Recall is the total number of predictions divided by the total number of true positives in the set. The testing result using the selected features is given in table 5. The selected features are used for classification. For classification of samples, we have employed the freely available Machine Learning package, WEKA [27]. Out of 300 images in the dataset, 208 were used for training and the remaining 92 for testing purposes.

Table 5: Results obtained by proposed method

Normal	100%
Malignant	92.78%
Benign	100%

The confusion matrix has been obtained from the testing part .In this case for example out of 97 actual malignant images 07 images was classified as normal. In case of benign and normal all images are correctly classified. The confusion matrix is given in Table 6

. Table 6: Confusion matrix

Actual	Predicted class		
	Benign	Malignant	Normal
Benign	104	0	0
Malignant	97	90	07
Normal	99	0	99

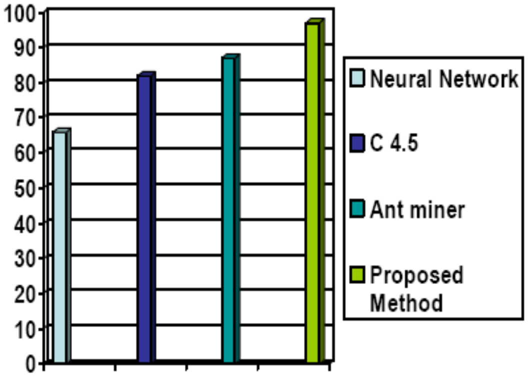


Fig. 5. Performance of the Classifier

7. Conclusion

Automated breast cancer detection has been studied for more than two decades Mammography is one of the best methods in breast cancer detection, but in some cases radiologists face difficulty in directing the tumors. We have described a comprehensive of methods in a uniform terminology, to define general properties and requirements of local techniques, to enable the readers to select the efficient method that is optimal for the specific application in detection of micro calcifications in mammogram images. In this paper, a new method for association rule mining is proposed. The main features of this method are that it only scans the transaction database once, it does not produce candidate jtemsets, and it adopts the Boolean vector "relational calculus" to discover frequent itemsets. In addition, it stores all transaction data in binary form, so it needs less memory space and can be applied to mining large databases.

The CAD mammography systems for micro calcifications detection have gone from crude tools in the research laboratory to commercial systems. Mammogram image analysis society database is standard test set but defining different standard test set (database) and better evaluation criteria are still very important. With some rigorous evaluations, and objective and fair comparison could determine the relative merit of competing algorithms and facilitate the development of better and robust systems. The methods like one presented in this paper could assist the medical staff and improve the accuracy of detection. Our method can reduce the computation cost of mammogram image analysis and can be applied to other image analysis applications. The algorithm uses simple statistical techniques in collaboration to develop a novel feature selection technique for medical image analysis.

Appendix

Appendixes, if needed, appear before the acknowledgment.

References

[1]. Majid AS, de Paredes ES, Doherty RD, Sharma N Salvador X. "Missed breast carcinoma: pitfalls and Pearls". Radiographics, pp.881-895, 2003.

[2]. Osmar R. Zaïane,M-L. Antonie, A. Coman "Mammography Classification by Association Rule based Classifier," MDM/KDD2002 International Workshop on Multimedia Data Mining ACM SIGKDD, pp.62-69,2002,

[3]. Xie Xuanyang, Gong Yuchang, Wan Shouhong, Li Xi ,"Computer Aided Detection of SARS Based on Radiographs Data Mining ", Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference Shanghai, China, pp7459 – 7462, 2005.

[4] C.Chen and G.Lee, "Image segmentation using multitiresolution wavelet analysis and Expectation Maximum(EM) algorithm for mammography" , International Journal of Imaging System and Technology, 8(5): pp491-504,1997.

[5] T.Wang and N.Karayaiannis, "Detection of microcalcification in digital mammograms using wavelets", IEEE Trans. Medical Imaging, 17(4):498-509, 1998.

[6] Jelena Bozek, Mario Mustra, Kresimir Delac, and Mislav Grgic "A Survey of Image Processing Algorithms in Digital

mammography”Grgic et al. (Eds.): Rec. Advan. in Mult. Sig. Process. and Commun., SCI 231, pp. 631–657,2009

[7]. Shuyan Wang, Mingquan Zhou and Guohua Geng, “Application of Fuzzy Cluster analysis for Medical Image Data Mining” Proceedings of the IEEE International Conference on Mechatronics & Automation Niagara Falls, Canada,pp. 36 – 41,July 2005.

[8]. R.Jensen, Qiang Shen, “Semantics Preserving Dimensionality Reduction: Rough and Fuzzy-Rough Based Approaches”, IEEE Transactions on Knowledge and Data Engineering, pp. 1457-1471, 2004.

[9]. I.Christiyanni et al ., “Fast detection of masses in computer aided mammography”, IEEE Signal processing Magazine, pp:54- 64,2000.

[10]. Walid Erray, and Hakim Hacid, “A New Cost Sensitive Decision Tree Method Application for Mammograms Classification” IJCSNS International Journal of Computer Science and Network Security, pp. 130-138, 2006.

[11]. Ying Liu, Dengsheng Zhang, Guojun Lu, Regionbased “image retrieval with high-level semantics using decision tree learning”, Pattern Recognition, 41, pp. 2554 – 2570, 2008.

[12]. Kemal Polat , Salih Gu¨nes, “A novel hybrid intelligent method based on C4.5 decision tree classifier and one-against-all approach for multi-class classification problems”, Expert Systems with Applications, Volume 36 Issue 2, pp.1587-1592, March, 2009, doi:10.1016/j.eswa.2007.11.051

[13]. Etta D. Pisano, Elodia B. Cole Bradley, M. Hemminger, Martin J. Yaffe, Stephen R. Aylward, Andrew D. A. Maidment, R. Eugene Johnston, Mark B. Williams,Loren T. Niklason, Emily F. Conant, Laurie L. Fajardo,Daniel B. Kopans, Marylee E. Brown, Stephen M. Pizer “Image Processing Algorithms for Digital Mammography: A Pictorial Essay” journal of Radio Graphics Volume 20,Number 5,sept.2000

[14] Pisano ED, Gatsonis C, Hendrick E et al. “Diagnostic performance of digital versus film mammography for breast-cancer screening”. NEngl J Med 2005; 353(17):1773-83.

[15] Wanga X, Wong BS, Guan TC. ‘Image enhancement for radiography inspection”. International Conference on Experimental Mechanics. 2004: 462-8.

[16]. D.Brazokovic and M.Nescovic, “Mammogram screening using multisolution based image segmentation”, International journal of pattern recognition and Artificial Intelligence, 7(6): pp.1437-1460, 1993

[17]. Dougherty J, Kohavi R, Sahami M. “Supervised and unsupervised discretization of continuous features”. In: Proceedings of the 12th international conference on machine learning.San Francisco:Morgan Kaufmann; pp 194–202, 1995.

[18]. Yvan Saeys, Thomas Abeel, Yves Van de Peer “Towards robust feature selection techniques”, www.bioinformatics.psb.ugent

[19] Gianluca Bontempi, Benjamin Haibe-Kains “Feature selection methods for mining bioinformatics data”, http://www.ulb.ac.be/di/mlg

[20]. Li Liu, Jian Wang and Kai He “Breast density classification using histogram moments of multiple resolution mammograms” Biomedical Engineering and Informatics (BMEI), 3rd International Conference, IEEE explore pp.146–149, DOI: November 2010, 10.1109/ BMEI.2010 .5639662,

[21]. Li Ke,Nannan Mu,Yan Kang Mass computer-aided diagnosis method in mammogram based on texture features, Biomedical Engineering and Informatics (BMEI), 3rd International Conference, IEEE Explore, pp.146 – 149, November 2010, DOI: 10.1109/ BMEI.2010.5639662,

[22] Azlindawaty Mohd Khuzi, R. Besar and W. M. D. Wan Zaki “Texture Features Selection for Masses Detection In Digital Mammogram” 4th Kuala Lumpur International Conference on Biomedical Engineering 2008 IFMBE Proceedings, 2008, Volume 21, Part 3, Part 8, 629-632, DOI: 10.1007/978-3-540-69139-6_157

[23] S.Lai,X.Li and W.Bischof “On techniques for detecting circumscribed masses in mammograms”, IEEE Trans on Medical Imaging , 8(4): pp. 377-386,1989.

[24]. Somol, P.Novovicova, J..Grim, J., Pudil, P.” Dynamic Oscillating Search Algorithm for Feature Selection” 19th International Conference on Pattern Recognition, 2008. ICPR 2008. pp.1-4 D.O.I10.1109/ICPR.2008.4761773

[25]. R. Kohavi and G. H. John. “Wrappers for feature subset selection”. Artif. Intell., 97(1-2):273–324, 1997.

[26] Deepa S. Deshpande “ASSOCIATION RULE MINING BASED ON IMAGE CONTENT” International Journal of Information Technology and Knowledge Management January-June 2011, Volume 4, No. 1, pp. 143-146

[27].Holmes, G., Donkin, A., Witten, I.H.: WEKA: a machine learning workbench. In: Proceedings Second Australia and New Zealand Conference on Intelligent Information Systems, Brisbane, Australia, pp. 357-361, 1994.

First Author: Aswini kumar mohanty has obtained his Bachelor of engineering in computer technology in 1991 from Mararthawada University in 1991 and M.Tech. in computer science from kalinga university in 2005.Currently he is pursuing his phd under SOA university in image mining.He has served in many engineering colleges and presently working as Associate professor in computer science department of Gandhi Engineering college, Bhubaneswar, odhisa, His area of interest is Computer Architecture, Operating system, Data mining, image mining. He has published more than 15 papers in national and international conferences and journals.

Second Author Sukanta Kumar swain has obtained his master degree in computer application (MCA) from Indira Gandghi Open University(IGNOU) in 2006 and currently pursuing his M.Tech degree in Computer Science under BPUT, Rourkela, Odisha and working as assistant professor in NIIS institute of Business Administration, Bhubaneswar, Odisha..

Third Author Pratap Kumar Champati obtained his bachelor of engineering in computer science in 1993 from Utkal University and has worked in industries as well as in academics. Currently he is working as assistant professor in ABIT engineering college, cuttack, Orissa and also pursuing M.Tech. in computer science under BPUT, Rourkela, Odhisa.

Fourth Author Dr. Saroj kumar lenka Passed his B.E. CSE in 1994 from Utkal Universty and M.tech in 2005.He obtained his PHDfrom Berhampur University in 2008 from deptt of comp. sc.Currently he is working as a professor in deptt. of CSE at MODI University, Rajstan. His area of research is image processing, data mining and computer architecture.

Study of the Master-Slave replication in a distributed database*

Kalonji Kalala Hercule[1]*, Mbuyi Mukendi Eugene[2], Boale Bomolo Paulin[3], Lilongo Bokaletumba Joel[4]

1.Department of Computer Science, Laboratoire d'information de Kinshasa, University of Kinshasa (unikin), Kinshasa, DR Congo.

2.Department of Computer Science, Laboratoire d'information de Kinshasa, University of Kinshasa (unikin), Kinshasa, DR Congo

3.Department of Computer Science, Laboratoire d'information de Kinshasa, University of Kinshasa (unikin), Kinshasa, DR Congo

4.Department of Computer Science, Laboratoire d'information de Kinshasa, University of Kinshasa (unikin), Kinshasa, DR Congo

Abstract

In a distributed database, data replication can be used to increase reliability, and availability of data. Updating a copy should be passed automatically to all its replicas. Generally, an update of data by the peer who has a new version involves not spread to those replicating this data. No form of consistency between replicas is guaranteed [13]. In this paper we propose the study "Replication Master-Slave", which is a way of replicating data used in a distributed database. We will do a brief overview on some principles of distributed databases. Then we will present the different types of replication, the value of using this mode of replication Master exclave. Then we will end up a banking application based on the replication Master-Slave on the Oracle 10g platform.

Keys words: *Distributed Database, replication, fragmentation*

I. Introduction

We consider a model of distributed database system similar to the model in [14-16]. Distributed database is a collection of multiple, logically interrelated database distributed over computer network [8]. In this model, a distributed database system consists of a set of data items residing at various sites. Sites can exchange information via messages transmitted on a communication network, which is assumed to be reliable. The distributed databases are born of the needs of the connected remote databases via a computer network while maintaining transparency in terms of localization, partitioning, and replication relative to the user. The replication of the databases is a well-known technique to improve the performances of the system and to avoid the failures of the sites [1-3-9-19]. Replication was commonly applied to distributed file systems to increase availability and fault tolerance [15]. Replication of databases is by definition the process of copying and maintaining database objects (such as tables, indexes) in databases that form a system of distributed database. Replication allows management of multiple copies that can diverge. Multiple copies may have different values at a given moment, but converging to the same values at another moment. Replication is an important mechanism because it allows organizations to provide their users with access to current data when they need it. It improves performance and increases data availability. Replication is a technique aimed at achieving goals toward both availability and fault-tolerance [22-23].

Replication object

A replication object is an object database such as a relationship, index, view, function or procedure that exists on multiple servers in a distributed database In a replication environment type, any changes to an object of this replication in a site, is reflected in all copies of that object in the other sites [3-16-9-21].

A replication group

In a replication environment, replication objects are managed using replication groups. A replication group is a collection of replication objects logically related [9-16]. The organization database connected to a replication group simplifies management of these objects.

Masters and slaves sites

A replication group can exist in multiple sites of replication. Replication environments support two basic types of sites: [24] masters sites and slaves sites. A replication group can be associated with one or more masters sites and slaves one or more sites. A same site [26] can be both a master site for a specified replication group and a slave site for another replication group, but one site can never be a master site and slave site for the same replication group. A Master Site controls a replication group and objects, maintaining a complete copy of all objects in the replication group and propagates any changes to group all copies located at Slaves sites. A slave site may contain all objects of a replication group or only a subset. [25]The slaves do not host sites, however only a snapshot of a replication group, as, for example, data of a relationship, captured at a given moment. Sites that contain snapshots are usually refreshed periodically to synchronize them with their masters sites. In the case of a replication environment with more than one master site, all masters sites communicate directly with one another to continually propagate data changes that occur in the replication groups.

Propagation strategy

When propagating an update to peer replicas, there are options of either synchronous or asynchronous propagation. In synchronous propagation, an update is sent to peer replicas before the result is returned to the client. In contrast, in asynchronous propagation the update is propagated after the result is returned to the client. Asynchronous propagation gives a better response time to the client. Replication protocols developed in the transactional model are examples of the synchronous propagation; a write operation has to be confirmed to have performed on all replicas before the result is returned. An asynchronous propagation example is demonstrated by the lazy replication architecture, where a set of updates performed at one replica is sent to other replicas in gossip messages from time to time.[17- 9-16 -13]

Asymmetric and symmetric replication

Asymmetrical replication is a technique for managing copies based on primary site only allowed to update and responsible for disseminating updates to the secondary copies. [18-16] The Symmetric replication allows simultaneous updates of all copies of different transactions. All sites can be updated.

Master-Slave property [24][25][26]

Duplicate data asynchronously are the property of a single site, the master or primary site, and may be updated by this site alone. Following the metaphor of the publication and subscription, the master site (the publisher) makes data available sites slaves (subscribers). Sites slaves subscribe to data held by the master site, which means they receive read-only copies on their own local systems. Any site may potentially be the master site data sets do not overlap, but only one site can update the master copy of a particular set of data, so that no dispute can not updated occur among sites. Here are some sample applications that illustrate the potential of this particular type of replication

The distribution and spread of centralized information

Dissemination of data describes an environment where data is updated in a central location and then replicated in read-only sites. For example, information on products, such as the list prices, could be maintained at the headquarters of a company and duplicated the form of read-only copies housed in subsidiaries (Fig 1a) [19-20]. The consolidation of data describes an environment where data is updated locally, and then met in a directory read-only in a specific location. This method gives ownership of data at each site and gives it certain autonomy. For example, details of properties maintained in each agency are duplicated in a copy of read-only preserved building on the site of the headquarters of the company. (Figure 1 b)

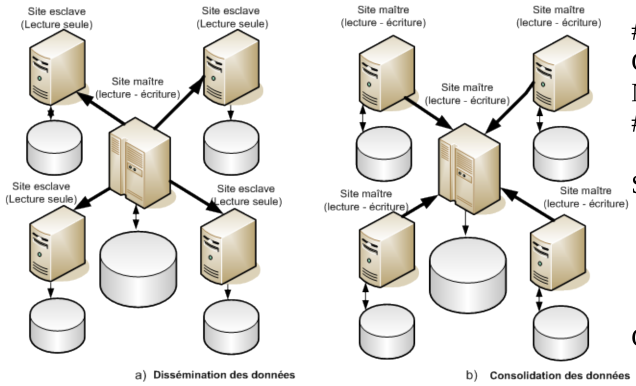


Fig. 1:Master Slave property (a.Data dissemination and b. Data consolidation

II.APPLICATION

The application consists of 3 sites:

- The central site Agency1: Who gathered all the data about customers, employees, accounts, agencies, operations that take place in other agencies (Agency2, Agency3). The central location also allows for creating and managing accounts for clients.

- The Agency2: Provides the ability to create and manage accounts for clients.

- The Agency2: Provides the ability to create and manage accounts for clients.

Representation of the conceptual model of the data

EMPLOYE (NumEmp,NomEmp,PrenomEmp,NumAg)
COMPTE (NumCompte,Numclient,TypeCompte,Somme,NumAg)
CLIENT (NumCli,Nom,Prenom,Age,AdresseAg)
AGENCE (NumAg,NomAg,AdresseAg)
OPERATION (NumOp,NumEmp,NumCompte,Montant,TypeCompt,DateOP)

Configuration

- Configuration du fichier listener.ora

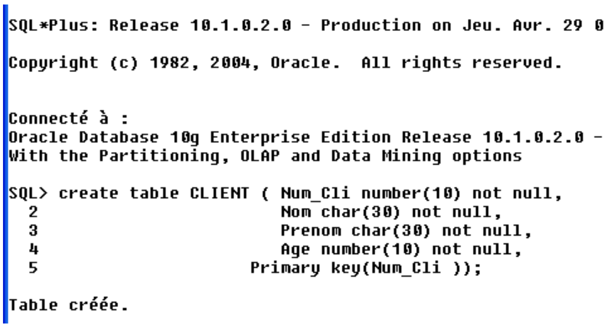
```
# listener.ora Network Configuration File: C:\oracle\product\10.1.0\Db_1\NETWORK\ADMIN\listener.ora
# Generated by Oracle configuration tools.

SID_LIST_LISTENER =
  (SID_LIST =
    (SID_DESC =
      (SID_NAME = PLSExtProc)
      (ORACLE_HOME = C:\oracle\product\10.1.0\Db_1)
      (PROGRAM = extproc)
    )
  )

LISTENER =
  (DESCRIPTION_LIST =
    (DESCRIPTION =
      (ADDRESS = (PROTOCOL = TCP)(HOST = localhost)(PORT = 1521))
    )
    (DESCRIPTION =
      (ADDRESS = (PROTOCOL = TCP)(HOST = DJO)(PORT = 1521))
    )
  )
```

- Configuration du fichier tnsnames.ora

```
# tnsnames.ora Network Configuration File: C:\Oracle\product\10.1.0\Client_1\network\admin\tnsnames.ora
# Generated by Oracle configuration tools.

AGENCE3 =
  (DESCRIPTION =
    (ADDRESS_LIST =
      (ADDRESS = (PROTOCOL = TCP)(HOST = DJO)(PORT = 1521))
    )
    (CONNECT_DATA =
      (SERVICE_NAME = AGENCE3)
    )
  )

AGENCE2 =
  (DESCRIPTION =
    (ADDRESS_LIST =
      (ADDRESS = (PROTOCOL = TCP)(HOST = DJO)(PORT = 1521))
    )
    (CONNECT_DATA =
      (SERVICE_NAME = AGENCE2)
    )
  )

AGENCE1 =
  (DESCRIPTION =
    (ADDRESS_LIST =
      (ADDRESS = (PROTOCOL = TCP)(HOST = DJO)(PORT = 1521))
    )
    (CONNECT_DATA =
      (SERVICE_NAME = AGENCE1)
    )
  )

DBPROD =
  (DESCRIPTION =
    (ADDRESS_LIST =
      (ADDRESS = (PROTOCOL = TCP)(HOST = DJO)(PORT = 1521))
    )
    (CONNECT_DATA =
      (SERVICE_NAME = DBPROD)
    )
  )

DBTEST =
  (DESCRIPTION =
    (ADDRESS_LIST =
      (ADDRESS = (PROTOCOL = TCP)(HOST = DJO)(PORT = 1521))
    )
    (CONNECT_DATA =
      (SERVICE_NAME = DBTEST)
    )
  )
```

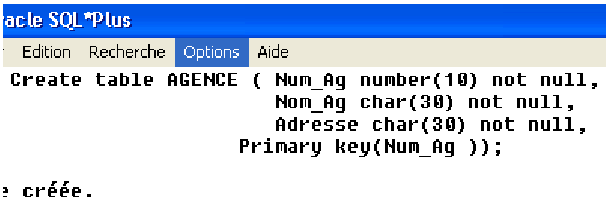
Creation des tables

Client table



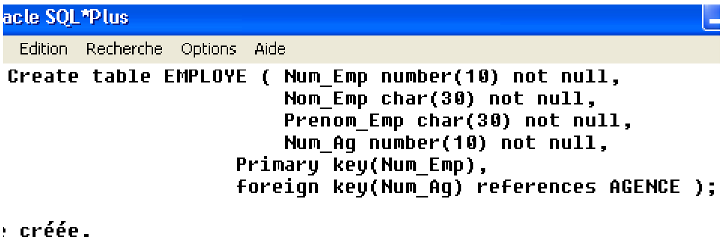
```
SQL*Plus: Release 10.1.0.2.0 - Production on Jeu. Avr. 29 0

Copyright (c) 1982, 2004, Oracle.  All rights reserved.


Connecté à :
Oracle Database 10g Enterprise Edition Release 10.1.0.2.0 -
With the Partitioning, OLAP and Data Mining options

SQL> create table CLIENT ( Num_Cli number(10) not null,
  2                        Nom char(30) not null,
  3                        Prenom char(30) not null,
  4                        Age number(10) not null,
  5                     Primary key(Num_Cli ));

Table créée.
```

Agence table



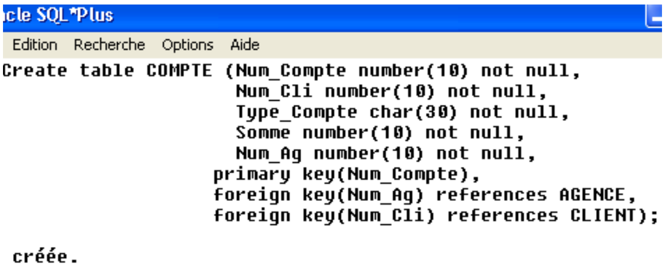
```
Create table AGENCE ( Num_Ag number(10) not null,
                      Nom_Ag char(30) not null,
                      Adresse char(30) not null,
                   Primary key(Num_Ag ));

: créée.
```

Employé table



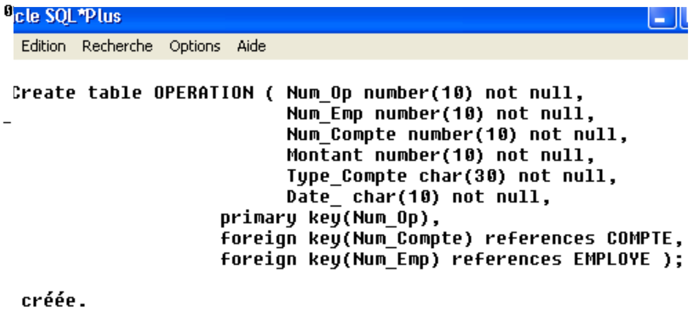
```
Create table EMPLOYE ( Num_Emp number(10) not null,
                       Nom_Emp char(30) not null,
                       Prenom_Emp char(30) not null,
                       Num_Ag number(10) not null,
                   Primary key(Num_Emp),
                   foreign key(Num_Ag) references AGENCE );

: créée.
```

Creation table Compte



```
Create table COMPTE (Num_Compte number(10) not null,
                     Num_Cli number(10) not null,
                     Type_Compte char(30) not null,
                     Somme number(10) not null,
                     Num_Ag number(10) not null,
                   primary key(Num_Compte),
                   foreign key(Num_Ag) references AGENCE,
                   foreign key(Num_Cli) references CLIENT);

 créée.
```

Creation table operation



```
Create table OPERATION ( Num_Op number(10) not null,
                         Num_Emp number(10) not null,
                         Num_Compte number(10) not null,
                         Montant number(10) not null,
                         Type_Compte char(30) not null,
                         Date_ char(10) not null,
                     primary key(Num_Op),
                     foreign key(Num_Compte) references COMPTE,
                     foreign key(Num_Emp) references EMPLOYE );

 créée.
```

Création de des liens et des synonymes

- Liens et synonymes sur AGENCE1

CREATE DATABASE LINK l1
CONNECT TO system
IDENTIFIED BY joe
USING 'AGENCE2';

```
CREATE SYNONYM CLIENT2 FOR
CLIENT@l1;
CREATE SYNONYM AGENCE2 FOR
AGENCE@l1;
CREATE SYNONYM EMPLOYE2 FOR
EMPLOYE@l1;
CREATE SYNONYM COMPTE2 FOR
COMPTE@l1;
CREATE SYNONYM OPERATION2 FOR
OPERATION@l1;

CREATE DATABASE LINK l3
CONNECT TO system
IDENTIFIED BY joe
USING 'AGENCE3';
CREATE SYNONYM CLIENT3 FOR
CLIENT@l3;
CREATE SYNONYM AGENCE3 FOR
AGENCE@l3;
CREATE SYNONYM EMPLOYE3 FOR
EMPLOYE@l3;
CREATE SYNONYM COMPTE3 FOR
COMPTE@l3;
CREATE SYNONYM OPERATION3 FOR
OPERATION@l3;
```

- Liens et synonymes sur AGENCE2

```
CREATE DATABASE LINK l2
CONNECT TO system
IDENTIFIED BY joe
USING 'AGENCE1';
CREATE SYNONYM CLIENT1 FOR
CLIENT@l2;
CREATE SYNONYM AGENCE1 FOR
AGENCE@l2;
CREATE SYNONYM EMPLOYE1 FOR
EMPLOYE@l2;
CREATE SYNONYM COMPTE1 FOR
COMPTE@l2;
CREATE SYNONYM OPERATION1 FOR
OPERATION@l2;

CREATE DATABASE LINK l3
CONNECT TO system
IDENTIFIED BY joe
USING 'AGENCE3';
CREATE SYNONYM CLIENT3 FOR
CLIENT@l3;
CREATE SYNONYM AGENCE3 FOR
AGENCE@l3;
CREATE SYNONYM EMPLOYE3 FOR
EMPLOYE@l3;
CREATE SYNONYM COMPTE3 FOR
COMPTE@l3;
CREATE SYNONYM OPERATION3 FOR
OPERATION@l3;
```

- Liens et synonymes sur AGENCE3

```
CREATE DATABASE LINK l3
CONNECT TO system
IDENTIFIED BY joe
USING 'AGENCE1';
CREATE SYNONYM CLIENT1 FOR
CLIENT@l3;
CREATE SYNONYM AGENCE1 FOR
AGENCE@l3;
CREATE SYNONYM EMPLOYE1 FOR
EMPLOYE@l3;
CREATE SYNONYM COMPTE1 FOR
COMPTE@l3;
CREATE SYNONYM OPERATION1 FOR
OPERATION@l3;


CREATE DATABASE LINK l2
CONNECT TO system
IDENTIFIED BY joe
USING 'AGENCE2';
CREATE SYNONYM CLIENT2 FOR
CLIENT@l2;
CREATE SYNONYM AGENCE2 FOR
AGENCE@l2;
CREATE SYNONYM EMPLOYE2 FOR
EMPLOYE@l2;
CREATE SYNONYM COMPTE2 FOR
COMPTE@l2;
CREATE SYNONYM OPERATION2 FOR
OPERATION@l2;
```

- **Data replication**

1. Using native command of sqlplus ie copy

- AGENCE 2

```
copy from system/joe@Agence1 to
system/joe@Agence2 Create CLIENT_T
(NUM_CLI,NOM,PRENOM,AGE) Using
SELECT * FROM CLIENT;
copy from system/joe@Agence1 to
system/joe@Agence2 Create COMPTE_T
(NUM_COMPTE,NUM_CLI,TYPE_CO
MPTE,SOMME,NUM_AG) Using
SELECT * FROM COMPTE;
copy from system/joe@Agence1 to
system/joe@Agence2 Create
OPERATION_T
(NUM_OP,NUM_EMP,NUM_COMPTE,
MONTANT,TYPE_COMPTE,DATE_)
Using SELECT * FROM OPERATION;
```

- AGENCE3

```
copy from system/joe@Agence1 to
system/joe@Agence3 Create CLIENT_T
```

```
(NUM_CLI,NOM,PRENOM,AGE) Using
SELECT * FROM CLIENT;

copy from system/joe@Agence1 to
system/joe@Agence3 Create COMPTE_T
(NUM_COMPTE,NUM_CLI,TYPE_COMPT
E,SOMME,NUM_AG) Using SELECT *
FROM COMPTE;

copy from system/joe@Agence1 to
system/joe@Agence3 Create OPERATION_T
(NUM_OP,NUM_EMP,NUM_COMPTE,MO
NTANT,TYPE_COMPTE,DATE_) Using
SELECT * FROM OPERATION;
```

The disadvantage of this method of replication: data can’t upgrade but you use replace command to change contain of tables.

2. Using the snapshot

Two types of snapshot can create: a simple or complex. A simple snapshot can’t contain a clause in list below : distinct, connect by , group by, join and operation set. The snapshot below extract master data and renew the operation three day after:

```
Create snapshot client_t
Tablespace data2
Storage (initial 100K next 100K PCINCREASE 0)
Refresh fast
Start with sysdate
Next sysdate + 3
As select * from client@link2;
```

Refresh fast use a snapshot log to upgrade a snapshot. This file locates on the same site of master table. In snapshot log, stored all modification intervene on master table. Thus, for each update, only the modifications which are sent, and not the whole of the data. On the other hand, a COMPLETE REFRESH is obligatory for the complex snapshots.

The snapshot log is to be created before the snapshot:

```
Create snapshot log on client
tablespace data
Storage (initial 10k next 10k pctincrease 0)
pctfree 5 pctused 90;
```

A traditional use of the snapshots in update is agencies. All the agencies store data concerning the accounts which they have created during the day. Each night, the data are poured in the national base which centralizes the data of the bank. Let us note that, the snapshots in update can generate conflicts. A release (trigger) saves the updates operated on the snapshot and transmits them to the main site at the time of the cooling of the snapshot.

3. Using materialized view

The materialized views are view whose contents are physically present (contrary to the traditional view which is select with delayed-action). The contents must be regularly refreshed. Several strategies are established.

3.1. Refresh complete

Create materialized view client_t refresh complete as select * from client@link2;

The request will be carried out on the distant table and the result will replace the current contents. Can be very greedy in resource and time for the bulky tables.

3.2. Refresh fast

Create materialized view client refresh fast as select * from client@link2;

The maitre records the operations made on the table; during a refresh of the replica only modifications since the last refresh one will be downloaded and applied locally.

A materialized view can bring several advantages to the level performances. According to the complexity of the request, we can fill it with changes, by means of the log of materialized views (MATERIALIZED VIEW LOG), instead of recreating it.

Contrary to snapshots, the materialized views can be directly used by the optimizer, to modify the path of execution of the requests.

III.CONCLUSION

In this paper, we presented how we can use the Master-Slave replication. We have shown the benefits of this type of replication, how it works, and ways to use depending on whether the updates are immediate or delayed. And we used an application that we designed to implement and demonstrate how this type of replication works. This application, we implemented the Oracle 10g. We also want to point out that this type of replication is suitable for applications in environments from the analysis of a system of decision support, or data from one or more databases are copied and loaded into a separate system of decision support for read-only analysis, in applications of the distribution and dissemination of information centralized, in consolidating information from remote and finally in applications of mobile computing or itinerant.

IV.REFERENCE

[1] G. Gardarin et O. Gardarin. "The Client-Server", Ed. Eyrolles 1996

[2] R. Bizoi. "Oracle 10g Administration", Tsoft Editeur, Ed. Eyrolles 2005

[3] S. Pelagatti. "Distributed database: Principes and systems", International edition. Copyright 1985.

[4] T. connolly and C. Beqq. "Systèmes de base de données – Approche pratique de la conception. De l'implémentation et de l'administration ",Edition Eyrolles, Reynald Goulet Juillet 2005.

[5] R. katz, e. wong. "Resolving Conflicts in Global Storage Design through Replication ", ACM Transactions on Database Systems, 1983

[6] M. Özsu, P. Valduriez. "Distributed database systems: where are we now?", IEEE Computer, 1991

[7] S. Upadhyaya, S. Lata. "Task allocation in Distributed computing VS distributed database systems : A Comparative study", IJCSNS International Journal of Computer Science and Network Security, Vol.8 N°3, 2008

[8] A. Singh K.S. Kahlon. "Non-replicated Dynamic Data Allocation in Distributed Database Systems", IJCSNS International Journal of Computer Science and Network Security, Vol.9, September 2009.

[9] J.E. Armendariz, J.R. Juarez1, J.R. Gonzaalez de Mendivil, F.D. Munoz "Correctness Criteria for Replicated Database Systems with Snapshot Isolation Replicas" Technical Report ITI-ITE-08/03, 2008.

[10] M. Özsu, P. Valduriez. "Distributed and Parallel Database Systems", 1991

[11] N. Conway, G. Sherry. "A Proposal for a Multi-Master Synchronous Replication" 2006.

[12] H. George, L. Fletcher, J. Bussche, D. Gucht, S. Vansummeren."Towards a Theory of Search Queries "ACM Transactions on Database Systems, 2010.

[13] E. Pacitti. “Réplication Asynchrone des données dans trois contextes”. Laboratoire d'Informatique de Nantes Atlantique. Université de Nantes, France, 2008.

[14] J. Wu, D. Manivannan and B. Thuraisingham "Necessary and sufficient conditions for transaction-consistent global checkpoint in a distributed database system", Information Sciences, Vol 179, pp 3659–3672, 2009.

[15] Q. Wenyu, L. Keqiu, B. Jiang, H. Shen, and D. Wu, "Dynamically Selecting Distribution Strategies for Web Documents According to Access Pattern", IJCSNS International Journal of Computer Science and Network Security, VOL.6 No.3A, March 2006.

[16] K. Segun, R. Hurson, V. Desai and A. Spink, "Transaction Management in a Mobile Data Access System", Parallel Computing Technologies Lecture Notes in Computer Science, Springer, Vol, pp.112-127 2127, 2001.

[17] W. Zhou, L. Wang and W. Jia, "An analysis of update ordering in distributed replication systems", Future Generation Computer Systems, Vol 20, pp 565–590, 2004.

[18] M. Masud, I. Kiringa, "Transaction processing in a peer to peer database network", Data & Knowledge Engineering, Vol 70, pp 307–334, 2011.

[19] I. Roussaki, M. Strimpakou, C. Pils, N. Kalatzis and N. Liampotis, "Optimising context data dissemination and storage in distributed pervasive computing systems", Pervasive and Mobile Computing, Vol6, pp,218-238, 2010.

[20] L. Wujuan, B. Veeravalli, "Design and analysis of an adaptive object replication algorithm in distributed network systems", Computer Communications, Vol31, pp 2005–2015, 2008.

[21] P. Beran, W. Mach, E. Schikuta and R. Vigne, "A Multi-Staged Blackboard Query Optimization Framework for World-Spanning Distributed Database Resources", Procedia Computer Science Vol 4, pp156–165, 2011.

[22] W. Zhou, L. Wang and W. Jia, "An analysis of update ordering in distributed replication systems", Future Generation Computer Systems Vol20, pp565–590, 2004.

[23] S. Khan, I. Ahmad, "Comparison and analysis of ten static heuristics-based Internet data replication techniques ", Journal of Parallel and Distributed Computing, Vol 68, pp113 – 136, 2008.

[24] F. Shrouf, M. Eshtay and K. Humaidan, "Performance Optimization for Mobile Agent Message Broadcast Model Using V-Agent", IJCSNS International Journal of Computer Science and Network Security, Vol8, August 2008. (maître Esclave)

[25] S. Gançarsk, H. Naacke, E. Pacitti and P. Valduriez, "The leganet system: Freshness-aware transaction routingin a database cluster", Information Systems, Vol32, pp320–343, 2007.

[26] F. Silva, H. Senger"Improving scalability of Bag-of-Tasks applications running on master–slave platforms", Parallel Computing, Vol35, pp57–71, 2009.

[26] F. Silva, H. Senger"Improving scalability of Bag-of-Tasks applications running on master–slave platforms", Parallel Computing, Vol35, pp57–71, 2009.

Automatic Keywords Extraction for Punjabi Language

Vishal Gupta[1] and Gurpreet Singh Lehal[2]

**[1] Assistant Professor Computer Science & Engineering, UIET,
Panjab University Chandigarh, UT, Pin Code-160014,India**

**[2] Professor Department of Computer Science,
Punjabi University Patiala, Punjab, Pin Code-147002, India**

Abstract

Automatic keywords extraction is the task to identify a small set of words, key phrases, keywords, or key segments from a document that can describe the meaning of the document. Keywords are useful tools as they give the shortest summary of the document. This paper concentrates on Automatic keywords extraction for Punjabi language text. It includes various phases like removing stop words, Identification of Punjabi nouns and noun stemming, Calculation of Term Frequency and Inverse Sentence Frequency (TF-ISF), Punjabi keywords as nouns with high TF-ISF score and title/headline feature for Punjabi text. The extracted keywords are very much helpful in automatic indexing, text summarization, information retrieval, classification, clustering, topic detection and tracking and web searches etc.

Keywords: *Punjabi keywords extraction, Keywords, Key phrases, TF-ISF*

1. Introduction

Keywords [2] are set of significant words in a document that give high-level description of the content for investigating readers and are useful tools for many purposes. They are used in academic articles to give an insight about the article to be presented. In a magazine, they give clue about the main idea about the article so that the readers can determine whether the article is in their area of interest. In a textbook they are useful for the readers to identify the main points in their mind about a particular section. They can also be used for search engines in order to return more precise results in shorter time. Since keywords describe the main points of a text, they can be used as a measure of similarity for text categorization. In summary, keywords are useful tools for scanning large amount of documents in short time. The extracted keywords are very much helpful in automatic indexing, text summarization, information retrieval, classification, clustering, topic detection and tracking [7] and web searches etc.

Despite the usefulness of the keywords, very few of the current documents include them. In fact many authors are not intended to extract keywords and do not denote them unless they are not explicitly instructed to do so. Extracting keywords manually is an extremely difficult and time consuming process, therefore it is almost impossible to extract keywords manually even for the articles published in a single conference. Therefore there is a need for automated process that extracts keywords from documents.

Existing methods about Automatic Keyword Extraction can be divided into four categories [6]:-

1) Simple Statistics Approach: These methods are simple and do not need the training data. The statistical information of the words can be used to identify the keywords in the document. Cohen uses NGram statistical information to automatically index the document. N-Gram is language and domain independent. Other statistical methods include word frequency, TF*IDF [1], word co-occurrence [4][8], etc.

2) Linguistics Approach [3]: These approaches use the linguistic features of the words mainly sentences and documents. The linguistic approach includes the lexical analysis, syntactic analysis discourse analysis and so on.

3) Machine Learning Approaches: Keyword Extraction can be seen as supervised learning, Machine Learning approach employs the extracted keywords from training documents to learn a model and applies the model to find keywords from new documents. This approach includes Naïve Bayes, Support Vector Machine, etc.

4) Other approaches: Other approaches about keyword extraction mainly combines the methods mentioned above or use some heuristic knowledge in the task of keyword extraction, such as the position, length, layout feature of words, html tags around of the words, etc.

Various extraction methods discussed are for single document but these can further applied to multiple documents as per their suitability [5].In Automatic Keywords extraction system for Punjabi language, we are using combination of statistical and linguistics approaches for Punjabi language.

2. Automatic Keywords Extraction for Punjabi Language

Various phases of automatic keywords extraction for Punjabi language are: 1)Removing stop words 2)Identification of Punjabi nouns and noun stemming 3)Calculation of Term Frequency and Inverse Sentence Frequency (TF-ISF) [1] 4) Punjabi keywords as nouns with high TF-ISF score 5)Title/Headline feature.

2.1 Removing Stop words from Punjabi text

Punjabi language Stop words are most frequently occurring words in Punjabi text like: ਦੇ dē, ਹੈ hai, ਨੂੰ nūṃ, ਨਾਲ nāl, ਤੋਂ tōṃ… etc. We have to eliminate these words from the original text otherwise, sentences containing them can get influence unnecessarily. We have made a list of Punjabi language stop words by creating a frequency list from a Punjabi corpus. Analysis of Punjabi corpus taken from popular Punjabi newspapers has been done. This corpus contains around 11.29 million words and 2.03 lakh unique words [9]. We manually analyzed these unique words and identified 615 stop words. In the corpus of 11.29 million words, the frequency count of these stop words is 5.267 million, which covers 46.64% of the corpus.
Some of the most commonly occurring stop words are displayed in Table1

Table 1. Punjabi language Stop words list

ਦੀ dī	ਤੋਂ tōṃ	ਹੋ hō	ਸਨ san
ਨੂੰ nūṃ	ਵੀ vī	ਉਹ uh	ਕੀਤੀ kītī
ਹੈ hai	ਕਿ ki	ਉਸ us	ਜਿਸ jis
ਨੇ nē	ਅਤੇ atē	ਕਰ kar	ਵਾਲੇ vālē
ਸੀ sī	ਹਨ han	ਪਰ par	ਕਰਕੇ karkē and so on……….

2.2 Identification of Punjabi nouns and Stemming

Input words are checked in Punjabi noun morph for possibility of nouns. Usually the words which are nouns with high TF-ISF scores are treated as keywords. Punjabi noun morph is having 74592 noun words in different forms. Examples of Punjabi nouns are shown in table2.

Table2. Punjabi Nouns list

ਪਹੀਆ pahīā	ਟੱਬਰ ṭabbar	ਸਿੰਗ siṅg
ਪਰਛਾਂਵਾਂ parchāṃvāṃ	ਘਰ ghar	ਹੱਥ hatth
ਪਲਾਟ palāṭ	ਖੰਭ khambh	ਆਰਾ ārā and so on…

The purpose of stemming [10][11] is to obtain the stem or radix of those words which are not found in dictionary. In Punjabi language noun stemming [12][13][14], an attempt is made to obtain stem or radix of a Punjabi word and then stem or radix is checked against Punjabi noun morph for the possibility of noun. An in depth analysis of corpus was made and the possible noun suffixes [16] were identified (Table 3) and the various rules for Punjabi word noun stemming have been generated. Results of Punjabi language noun stemmer [16] are given in table 4.

Table 3. Punjabi language nouns suffix list

◌ੀਆਂ īāṃ	ਿ◌ਆਂ iāṃ	◌ੂਆਂ ūāṃ	◌ਾਂ āṃ
◌ੀਏ īē	◌ੇ ē	◌ੀਓ Īō	ਿ◌ਓ iō
◌ੋ ō	◌ੀਆ īā	ਿ◌ਆ iā	◌ੀਂ īṃ
ਈ ī	◌ੋਂ ōṃ	ਵਾਂ vāṃ	ਿ◌ਉਂ iuṃ
ਈਆ īā	--	--	--

Table 4. Results of Punjabi language Noun stemmer

Punjabi Noun word	***Stem word***	***suffix***	***Punjabi Noun word***	***Stem word***	***suffix***
ਕਸਾਈਆ Kasāīā	ਕਸਾਈ kasāī	ਈਆ īā	ਮਾਹੀਆ māhīā	ਮਾਹੀ māhī	◌ੀਆ īā
ਘਰੋਂ gharōṃ	ਘਰ ghar	◌ੋਂ ōṃ	ਭਾਸ਼ਾਵਾਂ bhāshāvāṃ	ਭਾਸ਼ਾ bhāshā	ਵਾਂ vāṃ
ਲੜਕੀਆਂ laṛkīāṃ	ਲੜਕੀ laṛkī	◌ੀਆਂ īāṃ	ਆਗੂਆਂ āgūāṃ	ਆਗੂ āgū	◌ੂਆਂ ūāṃ
ਫੁੱਲਾਂ phullāṃ	ਫੁੱਲ phull	◌ਾਂ āṃ	ਲੜਕੋ laṛkō	ਲੜਕਾ laṛkā	◌ੋ ō

ਲੜਕਿਆਂ laṛkiāṃ	ਲੜਕਾ laṛkā	ਿਆਂ iāṃ	ਲੜਕੀਏ laṛkīē	ਲੜਕੀ laṛkī	ੀਏ īē
ਮੁੰਡੇ muṇḍē	ਮੁੰਡਾ muṇḍā	ੇ ē	ਲੜਕੀਓ laṛkīō	ਲੜਕੀ laṛkī	ੀਓ Īō
ਲੜਕਿਓ laṛkīō	ਲੜਕਾ laṛkā	ਿਓ iō	ਲੜਕਿਆ laṛkiā	ਲੜਕਾ laṛkā	ਿਆ iā
ਘਰੀਂ gharīṃ	ਘਰ ghar	ੀਂ īṃ	ਦਰਵਾਜਿਉਂ darvājiu ṃ	ਦਰਵਾਜਾ darvājā	ਿਉਂ iuṃ
ਪਰਾਂਦੇ parāndē	ਪਰਾਂਦਾ parāndā	ੇ ē	ਭਾਸ਼ਾਈ bhāshāī	ਭਾਸ਼ਾ bhāshā	ਈ ī

An In depth analysis of output of Punjabi noun stemmer has been done over 50 Punjabi documents of Punjabi news corpus. The efficiency of Punjabi language noun stemmer is 82.6%.

2.3 Calculation of Term Frequency-Inverse Sentence Frequency TF-ISF

The basic idea of TF-ISF [1] [15] score is to evaluate each word in terms of its distribution over the document. Indeed, It is obvious that words occurring in many sentences within a document may not be useful for topic segmentation purposes. It is used to evaluate the importance of a word within a document based on its frequency within a given sentence and its distribution across all the sentences within the document. The TF-ISF measure of a word w in a sentence s, denoted TF-ISF(w,s), is computed by:
TF-ISF(w,s)= TF(w,s)* ISF(w) where the term frequency TF(w,s) is the number of times that word w occurs in sentence s, and the inverse sentence frequency ISF(w) is given by the formula:
ISF(w) = log(|S|/ SF(w)) , where the sentence frequency SF(w) is the number of sentences in which the word w occurs. Top scored Punjabi words (Top 20%) with high value of TF-ISF scores are candidates for keywords from this phase.

2.4 Punjabi keywords as nouns with high TF-ISF score

In this phase, Punjabi keywords are extracted by performing intersection of noun keywords and keywords with high TF-ISF score (Top 20%) from previous phases. Those Punjabi words which are Punjabi nouns and with high TF-ISF scores are candidates for Punjabi Keywords.

2.5 Punjabi language Title/Headline Feature

Noun words appearing in title/headline (after removing stop words) are always more important. These words are treated as keywords. The union of these keywords and keywords coming from previous phase (noun words with high TF-ISF scores) are treated as final Punjabi keywords.

2.6 Algorithm for Punjabi Keywords Extraction

Step1:-From input Punjabi text remove stop words.

Step2:- Check the input words in Punjabi noun morph for the possibility of Punjabi nouns and if necessary, perform noun stemming.

Step3:- Calculate TF-ISF score of each remaining Punjabi word TF-ISF(w,s)= TF(w,s)* ISF(w) where TF(w,s) is the number of times that word w occurs in sentence s, and the inverse sentence frequency ISF(w) = log(|S|/ SF(w)) , where the sentence frequency SF(w) is the number of sentences in which the word w occurs.

Step4:-Top scored words (top 20%) with high TF-ISF scores are candidates for keywords from this phase.

Step5:- Perform intersection of Punjabi noun keywords and keywords with high TF-ISF scores. Punjabi nouns with high TF-ISF score are candidates of Punjabi keywords from this phase.

Step6:- Treat the noun words appearing in title/headlines as keywords (After removing stop words).

Step7:- The Union of keywords coming from step5 and step6 are final Punjabi keywords.

3. Results and Conclusions

An In depth analysis of output of Punjabi keyword extraction has been done over 50 Punjabi documents of Punjabi news corpus. The Precision, Recall and F-Score of Punjabi language keywords extraction are 80.4%, 90.6% and 85.2% respectively. 14.8% of errors are due to absence of certain Punjabi noun words in noun morph, dictionary mistakes, input text syntax mistakes and certain rules violations of noun stemming.

The Example Input Punjabi text is as follows:-

ਉਪ ਮੁੱਖ ਮੰਤਰੀ ਸੁਖਬੀਰ ਬਾਦਲ ਦੇ ਦੌਰੇ ਨੂੰ ਲੈ ਕੇ ਭੁੱਚੋ ਮੰਡੀ ਦੇ ਵਾਸੀਆਂ ਨੇ ਕੀਤੀ ਮੀਟਿੰਗ

ਭੁੱਚੋ ਮੰਡੀ, 8 ਜਨਵਰੀ (ਜਸਪਾਲ ਸਿੰਘ ਸਿੱਧੂ)- ਪੰਜਾਬ ਦੇ ਉਪ ਮੁੱਖ ਮੰਤਰੀ ਸ: ਸੁਖਬੀਰ ਸਿੰਘ ਬਾਦਲ ਅਤੇ ਲੋਕ ਸਭਾ ਮੈਂਬਰ ਬੀਬਾ ਹਰਸਿਮਰਤ ਕੌਰ ਬਾਦਲ ਦੇ 10 ਜਨਵਰੀ ਦੇ ਭੁੱਚੋ ਮੰਡੀ ਦੌਰੇ ਨੂੰ ਲੈ ਕੇ

ਮੰਡੀ ਨਿਵਾਸੀਆਂ ਨੇ ਨਗਰ ਕੌਂਸਲ ਭੁੱਚੋ ਮੰਡੀ ਦੇ ਦਫ਼ਤਰ ਵਿਖੇ ਵਿਸ਼ਾਲ ਮੀਟਿੰਗ ਕੀਤੀ। ਇਸ ਮੀਟਿੰਗ ਵਿਚ ਮੰਡੀ ਦੀਆਂ ਮੰਗਾਂ ਬਾਰੇ ਖੁੱਲ੍ਹ ਕੇ ਵਿਚਾਰ ਵਟਾਂਦਰਾ ਕੀਤਾ ਗਿਆ। ਪਤਾ ਲੱਗਾ ਹੈ ਕਿ ਸ: ਸੁਖਬੀਰ ਸਿੰਘ ਬਾਦਲ ਇਸ ਦਿਨ ਮੰਡੀ ਦੇ ਆਮ ਲੋਕਾਂ ਨੂੰ ਮਿਲਕੇ ਮੰਡੀ ਦੀਆਂ ਸਾਂਝੀਆਂ ਸਮੱਸਿਆਵਾਂ ਨੂੰ ਮੌਕੇ 'ਤੇ ਹੀ ਦੂਰ ਕਰਨ ਦਾ ਯਤਨ ਕਰਨਗੇ।

up mukkh mantrī sukhbīr bādal dē daurē nūṃ lai kē bhuccō maṇḍī dē vāsīāṃ nē kītī mīṭiṅg

bhuccō maṇḍī, 8 janvarī (jaspāl siṅgh siddhū)- pañjāb dē up mukkh mantrī sa: sukhbīr siṅgh bādal atē lōk sabhā maimbar bībā harsimrat kaur bādal dē 10 janvarī dē bhuccō maṇḍī daurē nūṃ lai kē maṇḍī nivāsīāṃ nē nagar kauṃsal bhuccō maṇḍī dē daftar vikhē vishāl mīṭiṅg kītī. is mīṭiṅg vic maṇḍī dīāṃ maṅgāṃ bārē khullh kē vicār vaṭāndrā kītā giā. patā laggā hai ki sa: sukhbīr siṅgh bādal is din maṇḍī dē ām lōkāṃ nūṃ milkē maṇḍī dīāṃ sāñjhīāṃ samssiāvāṃ nūṃ maukē 'tē hī dūr karan dā yatan karnagē.

Output Keywords of Punjabi Keywords extraction are as follows:-

ਮੁੱਖ	(mukkh)
ਮੰਤਰੀ	(mantrī)
ਮੰਡੀ	(maṇḍī)
ਵਾਸੀਆਂ	(vāsīāṃ)
ਮੀਟਿੰਗ	(mīṭiṅg)
ਪੰਜਾਬ	(pañjāb)
ਨਿਵਾਸੀਆਂ	(nivāsīāṃ)
ਮੰਗਾਂ	(maṅgāṃ)
ਵਿਚਾਰ	(vichār)
ਲੋਕਾਂ	(lōkāṃ)
ਸਮੱਸਿਆਵਾਂ	(samssiāvāṃ)
ਯਤਨ	(yatan)

Now in the conclusion, in this paper, we have discussed the Automatic Keywords extraction for Punjabi language. Punjabi nouns with high TF-ISF scores are candidates for Punjabi Keywords. Noun words appearing in the title/headlines are directly treated as keywords. The extracted keywords are very much helpful in automatic indexing, text summarization, information retrieval, classification, clustering, topic detection and tracking and web searches etc.

Most of the lexical resources used in pre-processing such as Punjabi Stop words list and Punjabi noun stemmer had to be developed from scratch as no work had been done in that direction. For developing these resources an in depth analysis of Punjabi corpus, Punjabi dictionary and Punjabi morph had to be carried out. This the first time some of these resources have been developed for Punjabi and they can be beneficial for developing other NLP applications in Punjabi.

References

[1] Neto, Joel al., “Document Clustering and Text Summarization”, In: Proc. of 4th Int. Conf. Practical Applications of Knowledge Discovery and Data Mining, London, 2000, pp. 41-55.

[2] David B. Bracewell and Fuji REN, “ Multilingual Single Document Keyword Extraction For Information Retrieval”, Proceedings of NLP-KE, 2005,pp. 517-522.

[3] Xinghua u and Bin Wu, “ Automatic Keyword Extraction Using Linguistics Features ”, Sixth IEEE International Conference on Data Mining(ICDMW’06), 2006.

[4] Chengzhi Zhang, “ Automatic Keyword Extraction From Documents Using Conditional Random Fields ”, Journal of Computational and Information Systems, 2008.

[5] www. wikipedia.org.

[6] Jasmeen Kaur and Vishal Gupta, “ Effective Approaches for Extraction of Keywords”, International Journal of Computer Science Issues IJSCI, Vol.7, Issue 6. Non 2010, pp. 144-148.

[7] Sungjick Lee, Han-joon Kim, “ News Keyword Extraction For Topic Tracking ”, Fourth International Conference on Networked Computing and Advanced Information Management, 2008 ,pp. 554-559.

[8] Y. Matsuo and M. Ishizuka, “ Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information ”, International journal on Artificial Intelligence Tools, vol.13, no.1, 2004,pp.157-169.

[9] Punjabi Unique word Corpus.

[10] Md. Zahurul Islam, Md. Nizam Uddin and Mumit Khan, A light weight stemmer for Bengali and its Use in spelling Checker”, Proc. 1st Intl. Conf. on Digital Comm. And Computer Applications (DCCA 2007), Irbid, Jordan, March 2007,pp.19-23.

[11] Praveen Kumar, Shrikant Kashyap, Ankush Mittal and Sumit Gupta, “A query answering system for E-learning Hindi documents”, South Asian Language Review, VOL.XIII, Nos 1&2, January-June, 2003.

[12] Mandeep Singh Gill, G.S. Lehal and S.S. Joshi, “Part of Speech Tagging for Grammar Checking of Punjabi”, The Linguistic Jornal Volume 4 Issue 1, 2009, pp.6-21

[13] www.advancedcentrepunjabi.org/punjabi_mor_ana.asp

[14] Ananthakrishnan Ramanathan and Durgesh Rao, “ALightweight Stemmer for Hindi”, Workshop on Computational Linguistics for South-Asian Languages, EACL, 2003.

[15] Rasim M. Alguliev and Ramiz M. Aliguliyev,“ Effective Summarization Method of Text Documents ”, Proceedings of International Conference on Web Intelligence, IEEE, 2005.

[16] Vishal Gupta and Gurpreet Singh Lehal,” Preprocessing Phase of Punjabi Language Text Summarization”, International Conference on Information Systems for Indian Languages Communications in Computer and Information

Science ICISIL2011, Volume 139, Part2, Springer-Verlag Berlin Heidelberg, 2011, pp. 250-253.

First Author's Biodata

Vishal Gupta is Assistant Professor in Computer Science & Engineering at University Institute of Engineering & Technology, Panjab university Chandigarh. He has done MTech. in computer science & engineering from Punjabi University Patiala in 2005. He is among University toppers. He has done BTech. in Computer Science & Engineering from Govt. Engineering College Ferozepur in 2003. He is also pursuing his PhD in Computer Science & Engineering from University College of Engineering, Punjabi University Patiala, under the supervision of Dr. Gurpreet Singh Lehal. He has developed a number of research projects in field of natural language processing including synonyms detection, automatic question answering and text summarization etc.

Second Author's Biodata

Professor Gurpreet Singh Lehal received undergraduate degree in Mathematics in 1988 from Panjab University, Chandigarh, India, and Post Graduate degree in Computer Science in 1995 from Thapar Institute of Engineering & Technology, Patiala, India and Ph. D. degree in Computer Science from Punjabi University, Patiala, in 2002. He joined Thapar Corporate R&D Centre, Patiala, India, in 1988 and later in 1995 he joined Department of Computer Science at Punjabi University, Patiala. He is actively involved both in teaching and research. His current areas of research are- Natural Language Processing and Optical Character recognition. He has published more than 25 research papers in various international and national journals and refereed conferences. He has been actively involved in technical development of Punjabi and has to his credit the first Gurmukhi OCR, Punjabi word processor with spell checker and various transliteration software. He was the chief coordinator of the project "Resource Centre for Indian Language Technology Solutions-Punjabi", funded by the Ministry of Information Technology as well as the coordinator of the Special Assistance Programme (SAP-DRS) of the University Grants Commission (UGC), India. He was also awarded a research project by the International Development Research Centre (IDRC) Canada for Shahmukhi to Gurmukhi Transliteration Solution for Networking.

Knowledge Collaboration in Higher Educational Institutions in India : Charting a Knowledge Management Solution

Mamta Bhusry[1], **Jayanti Ranjan**[2]

[1] **Department of Computer Science, Ajay Kumar Garg Engineering College**
Ghaziabad – 201009, India

[2] **Information Management Systems Department, Institute of Management Technology**
Ghaziabad – 201001, India

Abstract

Knowledge management (KM) is an essential consideration in higher educational institutions (HEIs) to ensure that knowledge flows efficiently between the people and processes. A crucial aspect of KM in HEIs that has not been addressed adequately is the unstructured nature of knowledge management and varying degrees of conformance to KM mechanisms in the functional domains. The paper aims to propose a knowledge management framework for HEIs and evaluate the institutions for KM mechanisms in order to reiterate on the urgent need for knowledge management support in higher education.

The evaluation of the framework indicated the nascent nature of knowledge management in higher educational institutions in India. The evaluation also indicated that KM in HEIs is highly unstructured and occurs in disparate activities of the institutions and identified the potential domains for improvement based on the K-ASD framework.

The practical implications of KM initiatives in HEIs include the enhancement in the overall effectiveness and efficiency. A KM system should be integrated into the institution's processes and work environment

Keywords: Knowledge management, higher education, knowledge creation, knowledge encapsulation, knowledge structuring, knowledge, knowledge dissemination

1. Introduction

The last decade has experienced a manifold growth in higher education in India. With the increase in the number of institutions, competition has increased. The pressures of competition have compelled higher educational institutions to start thinking like businesses (Brown, Duguid, 2000). All educational institutions develop and use knowledge. The question is what value is added to the products and services they deliver by the effective use this knowledge asset (Milam, John, 2001). The institutions have to attune themselves to develop strategies for enhanced planning and development of processes and activities. This requires that institutions must be able to respond timely to the dynamic technologies and increasing demands of stakeholders (Nagad, Amin, 2006). For this, the knowledge in the organization needs to be identified, transformed, stored and disseminated effectively. This paves the way to discern the urgent need for knowledge management (KM) initiatives which is a key asset.

Knowledge Management is the management of organizational information and knowledge by applying skills, experience, innovation and intelligence. Wiig(1996) defines knowledge as "the insights, understandings and the practical know-how that we all possess". According to Nonaka(1998), Tiwana(2000) and Zack(1999), there are two types of knowledge – tacit and explicit. Tacit knowledge is the form of knowledge that is subconsciously understood and applied, difficult to articulate, developed from direct experiences and action and usually shared through highly interactive conversation and shared experiences. Explicit knowledge, on the other hand, is easy to articulate, capture and distribute in different formats. It is formal and systematic (Nakkiran, Sewry, 2002, pp.235-245). Essentially KM needs to ensure that the right knowledge gets to the right people at the right time, and to help people share and put knowledge into action in ways that strive to improve organizational performance (O'Dell, Grayson, 1998). According to Handzic(n.d.), a central task of KM research is to find the best ways to cultivate, nurture and exploit knowledge at individual, group and organizational levels.

The increasing needs of the stakeholders and pressures of competition require higher educational institutions to react in a proactive and efficacious manner. However the institutions are unable to respond at the required pace which results into a chasm between the “need” and the “availability”. It is important to identify this gap and make efforts towards the efficient management of the institutional knowledge.

In this paper the authors introduce a KM framework to explain the major elements of knowledge management and the flow of knowledge in the HEIs. The framework identifies three mechanisms for knowledge management in higher educational institutions : knowledge acquisition, knowledge transformation and knowledge dissemination. Further the authors have validated the need for knowledge management in higher educational institutions by evaluating the framework for the functional domains in the institutions. The findings reveal the nascent nature of KM in higher educational institutions and the urgent need to adopt IT based KM initiatives. The framework will encourage HEIs to focus their KM initiatives on performance outcomes and implement these initiatives in alignment with their organizational strategies. This will result in achievement of the desired performance outcomes.

2. Related Work

Significant work has been done in the area of KM in higher educational system and many new requirements have been proposed by different people in this field.

Kidwell, et al.(2000, pp. 28-33) discussed why KM is vital to higher education systems and how an institution wide approach to KM can lead to exponential improvements in knowledge sharing – both explicit and tacit and the subsequent surge benefits. The work deals with the benefits of various knowledge management applications on educational institution processes such as research, curriculum development, student and alumni services, administrative services and strategic planning.

Ranjan and Khalil (2007,pp. 15-25) have argued that in order to build and develop a robust and thriving knowledge environment the institutes need to look beyond technology and develop the overall culture of accessing, collaborating and managing knowledge.

Yeh (2005,pp.35-42) presented the KM multi-modeling framework to propose four organizational strategies for higher education – culture, leadership, technology and measurement and three academic KM strategies – individual, institutional and network.

Nagad and Amin (2006, pp.60-65) concluded that effective KM may require significant change in culture and value, organizational structures and reward systems. In order to apply KM, knowledge and expertise must be readily accessible, understandable and retrievable.

Sedziuviene, Vveinhardt, J.(2009, pp. 79-90) concluded that to create a KM system in higher educational institutions it is necessary to point out the valuable knowledge, to create a methodology for receiving, transforming and consolidating knowledge, to activate and optimize the process of knowledge formation, transmission and evaluation, to perform spread of knowledge among the staff and students, to constantly perform knowledge monitoring and make decisions accordingly and to generate new knowledge and new technologies for knowledge transmission.

Rowley (2000, pp. 325-333)in the study on KM in higher education said that KM challenges lie in the creation of a knowledge environment and the recognition of knowledge as intellectual capital. Effective KM in higher education requires significant change in the culture and values, organizational structures and reward systems.

This paper is motivated by the above related research to explore the knowledge management scenario with respect to higher educational institutions in India. The objective of this paper is to develop a KM framework that facilitates the institutions to capture, structure and disseminate the knowledge created in the organization so that it is readily available to everyone – anytime, anywhere.

3. Concerns and Priorities for KM in HEIs in India

Higher education in India is offered by a variety of institutions – Central Universities, Affiliation Universities, Private Universities, Deemed Universities, Vocational Universities, affiliated colleges and institutions and institutions of national eminence. The higher education system in India has become very complex due to the pressing aspirations of a developing and vibrant democracy. To meet this growing demand, while the number of universities and colleges have increased immensely, the quality of services offered by the HEIs has fallen short of the expectations. The factors contributing to the gap between the expectations and the actual are as listed –

- Lack of focused institutional planning
- Lack of research and consultancy
- Lack of commitment at all levels
- Lack of academic structure that promotes creativity and innovation
- Lack of innovative teaching and learning

processes

- Out dated curriculum due to lack of timely revision
- Inappropriate standard of services provided to students and alumni
- Near non-existence of academic-industry collaboration
- Low consistency in decision making
- Slow pace of process delivery

The quality of education being offered in higher education is a question being debated widely. With the growing cost of higher education, the pressures for producing industry ready professionals and competition for performance, the question has become especially pertinent for all stakeholders – students, faculty, industry and the policy makers.

HEIs in India are facing the pressures for enhanced performance for the reasons argued by Ashish and Arun (2006) and others –

1 Increasing competition among higher educational institutions
2 Growing awareness about alternate opportunities and value for money among the students and parents
3 Accountability to stakeholders and the accreditation and affiliating bodies
4 Increasing industry demands as employers for recruitments of graduates and post graduates
5 Industry expectations for industry-institution partnerships

In view of the pressures from the stakeholders and the present scenario in HEIs it becomes pertinent to look for solutions which will make an impact on the existing systems. A blend of KM and IT techniques can offer an appropriate tool to meet this challenge (Kumar and Kumar, 2005).

Large number of organizations have implemented KM principles and methods in their routine activities for enhanced collaborating of knowledge on inter and intra organizational platforms. However HEIs have not taken much interest in introducing KM approaches even though from the academic learning point of view KM by its nature is essential for HEIs (Ranjan and Khalil, 2007). Today HEIs behave like educational markets. They have to adjust themselves and develop strategies to respond rapidly to the increasing demands of stakeholders and market pressures.

A KM approach in HEIs is a conscious integration of all human resources and academic and administrative processes for the acquisition, structuring and sharing of institutional knowledge. Emphasis is required on sharing of knowledge at the institutional level and not the individual level (Ranjan and Khalil, 2007)

Higher educational institutions create knowledge during their academic and administrative processes. Knowledge is created at various levels in different forms and is required at each level in a different form. The processes of teaching, examination, evaluation, admissions, counseling, training and placement and research and consultancy result in numerous beneficial experiences and studies which may be defined as knowledge in the context of higher educational institutions (Ranjan and Khalil, 2007). A crucial aspect that has not been addressed adequately in higher educational institutions is the extent to which KM mechanisms have been implemented to share the institutional knowledge across the important functions and stakeholders in the institution. Efforts are needed to share the institutional knowledge in a cross functional and collaborative manner. KM in HEIs requires management of knowledge as an asset to recognize its value to the institution. This can be facilitated via an IT based KM paradigm that blends the KM processes with IT tools to improve the efficiency and effectiveness of HEIs.

4. Knnowledge Acquisition, Structuring and Dissemination (K-ASD): A Framework for KM in Higher Educational Institutions

The authors propose a framework that identifies the mechanisms for knowledge management in higher educational institutions. The framework focuses on the integrated acquisition of knowledge from all aspects of the organization and its deployment in the form as required by the stakeholders. The framework consists of three levels as shown in the figure.

Knowledge Acquisition – It is the mechanism through which knowledge is gathered and stored from the members of the institution and other resources (Schwartz, et al., 2000). According to Tiwana (2000), knowledge acquisition is the development and creation of insights, skills and relationships supported by information technology. Knowledge acquisition consists of codifying explicit knowledge, modulating tacit knowledge to explicit knowledge and codifying the explicit knowledge and acquiring tacit knowledge in the form of explicit meta knowledge i.e. knowledge about knowledge. The explicit meta knowledge about tacit knowledge contains information about “who knows what” and about how to contact the experts. The purpose of codification is to make it easy to organize, locate, share, store and use knowledge (Davenport and Prusak, 1998).

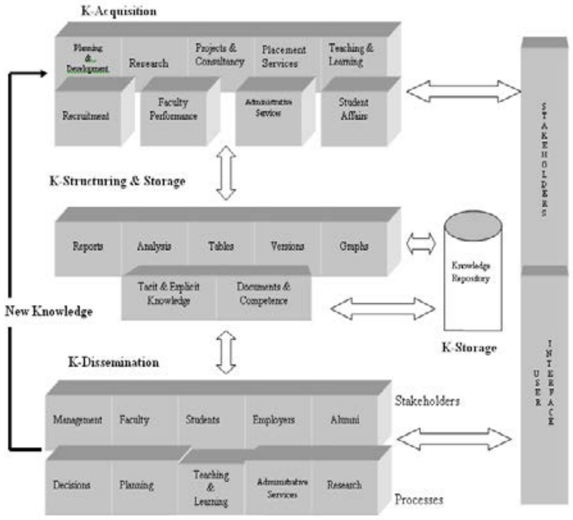


Fig. 1 K-ASD Framework

Knowledge Structuring and Storage– Knowledge may be created and acquired, but if not organized and structured, the organization will not be able to take action on that knowledge or actualize all of its potential value (O'Leary, n.d.). This will result into limited use of the institutional knowledge and its impact on the institution. Knowledge has to be structured into a form which can be used directly in the institutional processes and functions to "fit" into the institution's way of doing things. The form to which knowledge is converted is critical to the ability to use the knowledge. Under the circumstances institutions need to put the knowledge into specific forms viz. documents, databases, pictures, graphs, rules, case based reasoning (CBR) and frequently asked questions (FAQs). It includes organizing, indexing and formatting the acquired knowledge (Schwartz, et al., 2000) for reuse and leverage it in other ways and make it broadly available in the institution.

The knowledge is transformed into appropriate form as used and sought for by the stakeholders and stored in knowledge bases called knowledge repositories. A knowledge repository is a structured collection of the knowledge generated in an organization. This also encompasses the documents generated and the tacit knowledge available with the stakeholders, explicitly codified. The value of organizational knowledge increases when it is available in storage repositories for present and future use (Jasimuddin, 2005). The knowledge repository ensures the availability of related knowledge quickly and efficiently at the same place. A knowledge repository will contain the knowledge itself and information on knowledge. According to Natali and Falbo (n.d.), the primary requirement of the knowledge repository is to prevent the loss of knowledge and enhance accessibility to organizational knowledge in the form of a centralized well structured resource.

Knowledge Dissemination – The stored knowledge, if not transferred for further use within the organization, leads to wastage of organizational resources (Jasimuddin, 2005). According to Schwartz, et al.(2000), knowledge dissemination constitutes retrieval of the relevant knowledge for use at the right time. It supports the flow of knowledge in the institution. Knowledge dissemination refers to the transfer and deployment of knowledge to the points of use – people, practices, technology, products and services - through training, education and automated knowledge based systems. Knowledge dissemination can be pull based or push based as either the user can search for the required knowledge or the knowledge management system can offer knowledge that seems relevant for the user's task (Abecker, et al., 1998). Proactive knowledge dissemination becomes particularly important when users are not motivated to look for information, are too busy or unaware that relevant knowledge exists or are ignorant of the need for information in the first place (Natali and Falbo, n.d.). Dissemination of knowledge, active or passive, is not sufficient. The use of knowledge obtained from the organizations collective memory repository becomes quite involved. Activities such as proactive access, personalization and in particular tight integration with the user task play a crucial role for the effective reuse/application of knowledge. The responsibility of contextual interpretation and evaluation of the knowledge lies with the user. The knowledge is utilized and leveraged to act effectively for viability and success.

5. Evaluating HEIs for Knowledge Management using the K-ASD Framework

The authors performed an evaluation process to establish the validity of the framework based on the knowledge needs of higher educational institutions and the missing links that exist due to the lack of KM initiatives.

5.1 Design of variables and checklist of factors/determinants for KM

The authors identified the functional domains in the institutions and the determinants or factors that impact the effectiveness of KM in these domains via an interview and group discussion based study. Inputs were also gathered from work already accomplished in the field of KM in higher education (Ashish and Arun, 2006, Ranjan and

Khalil, 2007). The determinants identified were used as variables to evaluate the proposed framework for the existing scenario in HEIs. The domains and the determinants were distinguished on the basis of information collected during group and individual interviews with the faculty, heads of departments, deans and staff and observations of the procedures and processes. The data collected was analysed using the content analysis technique. Content analysis consists of analyzing the contents of documentary materials(books, magazines, newspapers) and verbal materials (interviews, group discussions) for the identification of certain characteristics that can be measured or counted (Kothari, 2010). The domains and determinants were further validated by two independent reviewers who were familiar with the higher education system in India and understood the objectives of the present work and the concepts underlying KM implementation in HEIs. The content analysis resulted in the identification of the activity domains in higher educational institutions and the determinants for KM intervention in these domains (Refer appendix 1).

5.2 Research Methodology

A survey based study was conducted as a preferred method for the research (Judd, et al., 1991). Based on the domains and determinants for KM intervention, the authors developed a checklist to evaluate the proposed framework in higher educational institutions. The objective of the study was to check the validity of the framework for the mechanisms in HEIs and establish the support for structured knowledge management. The higher educational institutions were chosen in the National Capital Region of Delhi, the names of which have not been disclosed.

To conduct the survey, the checklist was distributed to the respondents partly by mail and partly in person. An explanatory note on knowledge management and its benefits, role of for KM in HEIs, and the implications of IT based KM initiatives was distributed along with the checklist. It also outlined the context and the meaning of participation in the survey and the proposed uses of the data collected from the survey. Follow up telephone calls and e-mails were made to remind the respondents that the survey should be completed in order to maximize the response rates. It took about one month to complete the survey wherein 152 responses were received out of a total of 450 forms distributed. The response rate of the survey was 33.77%. The selection of the respondents was done very carefully keeping in mind the nature of the institutions, academic qualifications, designations and professional experience. The respondents were chosen from universities, engineering colleges and business schools.

In answering the questionnaire a determinant in the checklist was marked “YES” if compliance to knowledge capture, knowledge structuring and/or knowledge dissemination existed, else it was marked “NO”. The responses were encoded, entered into the computer (Excel Worksheets) and results computed in the following ways –

a) For each determinant the number of “YES” responses were added to find the percentage of compliance to knowledge management for the KM mechanisms (Appendix 2).

b) The average score and the percentage of compliance to KM for the functional domains was computed for each mechanism of the proposed framework (Appendix 4)

c) The average score and the percentage of compliance for sample HEIs was computed for the mechanisms of the KM framework (Appendix 3).

5.3 Empirical Evaluation

The results of the evaluation of data collected are given in appendix 2 to 4. For the functional domains of the higher educational institutions, the compliance to KM mechanisms exhibited a downward trend from knowledge acquisition to knowledge structuring and storage to knowledge dissemination (Refer Appendix 4). Such downward drift clearly implies that though knowledge in HEIs is captured and acquired to an extent of 39.77%, the focus on knowledge structuring and storage and knowledge dissemination is much less, a maximum of 36.06% and 32.61% respectively (Refer Appendix 3). These results indicate that the favourability for KM performance is poorer in the knowledge structuring mechanism as compared to knowledge acquisition and even poorer in the knowledge dissemination mechanism. The implication is that though higher educational institutions are acquiring, capturing and storing the institutional knowledge in different forms, there is least support for knowledge dissemination to make it easily available across the organization.

The mean and variance values (Appendix 4) for knowledge acquisition, knowledge structuring and storage and knowledge dissemination indicate that though the mean percentage of the conformance to knowledge management is highest for knowledge acquisition mechanism and least for knowledge dissemination mechanism, the variance shows a downward trend from knowledge acquisition to knowledge dissemination. Such sliding drift in dispersion from the mean value implies that KM practice is most ad hoc and non- structured in the knowledge dissemination phase of knowledge management.

Reference to appendix 2, the average score for knowledge acquisition in the functional domains varied from 29.32% (industrial projects and consultancy) to 48.02% (institutional administrative services), that for knowledge structuring and storage varied from 26.03% (industrial projects and consultancy) to 43.55% (institutional administrative services) and for knowledge dissemination it varied from 21.33% (industrial projects and consultancy) to 40.00% (institutional administrative services and faculty recruitment process both). The implication is that KM in HEIs exists in the form of a series of unrelated knowledge based activities which is not sufficient. The KM activities and practices should not occur in disparate pockets of the institutions; an organization needs to demonstrate these practices and activities throughout the organization, across all levels and groups in order to be a KM-smart organization (O'Leary, n.d.).

The lack of support for KM, ad hoc nature of KM and the downward trend of KM mechanisms in HEIs may be attributed to the reasons namely -

1. Lack of interest and confidence in using others' knowledge
2. Fear of losing importance by dispensing with one's knowledge
3. Silos mentality and lack of co-operation among employees
4. Lack of time
5. Lack of infrastructure (push and pull mechanisms)
6. Lack of organizational strategy for knowledge management
7. Lack of incentives to participate/collaborate for knowledge sharing

The findings indicate that knowledge management is a nascent concept in HEIs with ad hoc mechanisms and no defined structure. Most knowledge management takes place as part of routine processes of information storage in the institutions. There is indispensable need for effort and investment in both social and technical infrastructure in order to fully facilitate KM processes in HEIs. The KM mechanisms can be accomplished using substantial human effort supported by IT along with the overall culture of knowledge sharing.

6. Research and Practical Implications

The study has important implications for research and practice. KM in HEIs is at an emerging stage and there is scope for tremendous improvement in this area. KM involves interactions among people and processes across functions and domains influencing the knowledge sharing culture. The study supports the consideration of a holistic view of KM that integrates the interplay between the departments and the sections in the HEIs. Work on KM practices in HEIs is on the way towards better success, however more needs to be learnt and done about the effectiveness of KM initiatives in all respects in the institutions.

The practical implications of IT based KM initiatives in HEIs imply that the framework should be useful to the institutions in many ways namely –

- Enhanced ability to develop strategic plans
- Enhanced quality of programs and processes by identifying and leveraging best practices
- Enhanced ability to monitor and sustain ongoing change (Petrides, 2004)
- Enhanced faculty development efforts
- Improved teaching learning processes
- Improved effectiveness and efficiency of administrative services
- Improved sharing of internal and external information to minimize redundant efforts
- Reduced effort and turnaround time for actions
- Reduced operational costs

To gain user acceptance, a knowledge management system must be integrated into the organization's process, allowing to collect and store relevant knowledge as it is generated in the processes and functions of the organization (Natali and Falbo, n.d.). Consequently it should also be integrated to the existing work environment (Abecker, et al., 1998).

7. Discussion and Conclusion

Knowledge management is a crucial consideration in higher educational institutions to ensure that knowledge flows efficiently to the functionaries, students and other stakeholders. In this paper the authors have contributed a knowledge management framework for higher educational institutions characterized by a set of factors that impact the IT based KM initiatives in HEIs. Owing to the diversity in the functional domains of higher educational institutions and the determinants in each domain, the study has distilled only the more relevant domains and determinants for the evaluation of the framework. The framework can be utilised to cover a wider aspect of higher educational institutions. IT based KM intervention was found to be low in all the functional domains.

The authors argue that KM initiatives in higher educational institutions be used as part of institutional strategies to identify the knowledge needs and wants of stakeholders, design services to fulfill the knowledge needs and quantitatively and qualitatively measure the effectiveness of the knowledge management initiatives. An institution needs to have consistent and well defined expectations and opportunities for sharing information organization wide (Petrides, 2004).

The proposed KM framework can be used as a guide to develop institutional knowledge management models based on the institutional goals and objectives, functional domains and the determinants that will impact KM initiatives. With respect to IT, the framework is a significant knowledge enabler. This can be explained in terms of the potential of IT infrastructure in facilitating KM processes by providing a platform for knowledge storage and sharing. Information technology can be successfully used to facilitate knowledge acquisition and dissemination, knowledge storage in the form of a knowledge repository accessible to all the members of the organization, supporting collaboration among employees and fostering centered, real time, integrated systems. Thus IT can play an important role in advancing the institutional knowledge.

The implementation of a knowledge management framework in higher educational institutions will provide the stakeholders with opportunities of cross functional, inter and intra organizational knowledge sharing, collaborative problem solving, enhanced decision making, shorter development cycles and building of the competitive advantage. At the same time the implementation will be impeded by factors contributing to integration of various processes and cycles, conversion to services, adaptability threats, lack of commitment and barriers to knowledge sharing attributed by institutional hierarchy, geographical barriers and human nature.

Appendix 1: Functional Domains and Determinants for KM

	Planning and Development	**Institutional Research**
D1	Goals, objectives, vision, mission, targets and quality policy	In house publications
D2	Plans and policies	Research areas
D3	Reports by review committees and accreditation bodies	E-journals
D4	Competitor data	Latest trends in research
D5	Data related to assessment of procedures and processes	List of journals
D6		Research grants and facilities
	Industrial Projects and Consultancy	**Placement services**
D1	Project synopsis/proposals	Company data (salary packages, turnover, job profiles, promotion policies
D2	Consultancy areas of faculty	Industry trends
D3	Cost and time estimates	Approved procedures and processes
D4	Data on project failures	Top recruiters
D5	Project team structures used	Feedback from companies
D6	Deployment of resources to project teams	Nature of interview sessions
D7	Clients / customers	Alumni data
	Faculty recruitment process	**Institutional administrative services**
D1	Areas in which faculty is generally surplus / deficient	Procedures and formats of all forms and reports
D2	Faculty cadre ratio	Copy of schedules
D3	Curriculum	Rules and regulations
D4	Reasons for faculty mobility	HR policies for training and promotions
D5	Administrative responsibilities expected from faculty	Minutes of meetings
	Institutional teaching and learning process	**Performance Evaluation of the faculty**
D1	Teaching material	Results
D2	Course plans	Publications
D3	Curriculum	Industrial Consultancy
D4	Question banks, assignments and case studies	Student Project work
D5	Typical problems faced by students	Student feedback
D7	Frequently asked	Seminars and

	Questions (FAQs)	conferences organized by the faculty
D8	Effective teaching methodologies used by faculty for specific topics	Seminars, workshops and conferences attended by the faculty
D9	Related research	Administrative responsibility
D 10	Related projects	Personal Skills
D 11	Industry interfaces	Initiatives for self improvement and career development
	Student Affairs	
D1	Updated database of institutional resources, policies and procedures related to admissions, examinations, fees, financial aids, student counseling facilities, library etc.	
D2	A portal for placement facilities hosting information employers with contact details, package offered, job profile etc	
D3	Updated database of co-curricular and extra curricular activities and resources	
D4	Frequent problems encountered by students and their solutions	

Appendix 2: Conformance to KM for Determinants in HEI Functional Domains

(KA – Knowledge Acquisition, KS – Knowledge Structuring & Storage, KD – Knowledge Dissemination)

		KA %	KS %	KD %
Institutional Planning and Development	D1	41.45	38.82	31.58
	D2	42.76	39.47	32.89
	D3	38.16	34.21	34.21
	D4	30.92	29.61	24.34
	D5	32.24	28.29	25.66
Institutional Research	D1	44.74	36.18	38.16
	D2	45.39	33.55	34.87
	D3	46.71	42.76	36.18
	D4	36.18	30.92	28.29
	D5	46.05	44.74	39.47
	D6	41.45	38.16	32.89
Industrial Projects and Consultancy	D1	36.18	28.29	25
	D2	34.21	31.58	24.34
	D3	26.32	20.39	17.76
	D4	25.66	21.05	14.47
	D5	26.32	23.03	17.76
	D6	26.97	23.68	18.42
	D7	29.61	34.21	31.58
Placement	D1	47.37	44.74	41.45
services	D2	45.39	42.76	42.76
	D3	48.03	44.74	42.76
	D4	45.39	42.11	40.79
	D5	21.71	23.68	17.11
	D6	18.42	17.11	16.45
	D7	48.03	36.18	32.89
Institutional teaching and learning process	D1	47.37	36.18	28.29
	D2	47.37	44.74	43.42
	D3	47.37	46.05	44.74
	D4	41.45	29.61	28.29
	D5	29.61	23.03	16.45
	D6	28.29	23.03	18.42
	D7	26.32	16.45	14.47
	D8	34.21	30.92	28.95
	D9	30.92	25	21.05
	D10	26.97	21.71	18.42
	D11	23.03	18.42	17.11
Faculty recruitment process	D1	49.34	48.03	44.74
	D2	48.03	46.05	44.74
	D3	47.37	46.71	42.76
	D4	29.61	28.29	23.03
	D5	49.34	48.03	44.74
Performance Evaluation of the faculty	D1	48.03	45.39	38.82
	D2	47.37	44.74	36.84
	D3	48.68	45.39	40.13
	D4	48.03	45.39	42.76
	D5	29.61	25	19.08
	D6	44.74	43.42	41.45
	D7	47.37	44.74	41.45
	D8	46.05	44.74	43.42
	D9	48.68	47.37	44.08
	D10	44.74	40.13	40.13
	D11	48.03	44.74	40.13
Institutional administrative services	D1	48.03	46.05	41.45
	D2	48.03	39.47	38.16
	D3	48.68	44.74	37.5
	D4	46.05	40.79	39.47
	D5	49.34	46.71	43.42
Student Affairs	D1	49.34	48.03	47.37
	D2	40.13	39.47	33.55
	D3	33.55	32.89	30.92
	D4	29.61	28.29	27.63

Appendix 3 : Conformance to Knowledge Management the K-ASD Framework

Phases of KM Framework	**No. of Deter-minants**	**% of Conformance to KM**
Knowledge Acquisition	61	39.77%
Knowledge Structuring	61	36.06%

Knowledge Dissemination	61	32.61%

Appendix 4 Conformance to Knowledge Management Mechanisms in HEI Functional Domains

Functional Domain	Percentage of Conformance to Knowledge Management			
	No. of Deter-minants	KA %	KS %	KD %
Institutional Planning and Development	5	37.11	34.08	29.74
Institutional Research	6	43.42	37.72	34.98
Industrial Projects and Consultancy	7	29.32	26.03	21.33
Placement Services	7	39.19	35.90	33.46
Institutional Teaching and Learning Process	11	34.81	28.65	25.42
Faculty recruitment Process	5	44.74	43.42	40.00
Performance Evaluation of Faculty	11	45.57	42.82	38.94
Institutional Administrative Services	5	48.02	43.55	40.00
Student Affairs	4	38.16	37.17	34.87
MEAN		39.77	36.06	32.61
VARIANCE		86.802	91.584	97.768

References

[1] Abecker, A., Bernardi, A., Hinkelman, K., 1998, “Towards a Technology for Organizational Memories”, IEEE Intelligent Systems, Vol. 13, No. 3, pp. 40-48

[2] Ashish, Arun, 2006, “IT Based KM in Indian Higher Education System: Addressing Quality Concerns and Setting the Priorities Right”, Journal of Knowledge Management Practice, vol.7, No.3

[3] Brown, J.S., Duguid, P., 2000, “The Social Life of Information” Harvard Business School Press

[4] Davenport, T.H., Prusak, L., 1998”, “Working Knowledge : How Organizations Manage What They Know”, Harvard Business School Press

[5] Huveida, R., Shams, G., Hooshmand, A., 2008, “Knowledge Management Practices in Higher Education Institutions : A Different Approach”, IEEE 978-1-4244-2917-2,pp. 695-702

[6] Jasimuddin, S.M., 2005, “An Integration of Knowledge Transfer and Knowledge Storage : An Holistic approach”, GESTS International Transactions of Computer Science and Engineering, Vol. 18, No.1 pp.37-48

[7] Judd, C.M., Smith, E.R., Kidder, L.H., 1991, Research Methods in Social Relations, 6th ed., Harcourt Brace Jovanovich College Publishers

[8] Kidwell,J.,J., Vander Linde, K.M., Johnson, S.L., 2000, “Applying Corporate Knowledge Management Practices in Higher Education”, Educause Quaterly, pp. 28-33

[9] Kothari, C.R., 2010, Research Methodology Methods and Techniques, 4th ed., New Age International Publishers

[10] Kumar, A., Kumar, A., 2005, “IT Based Knowledge Management for Institutions of Higher Education- A Need”, University News, Vol. 43, No. 30, July 25-31, pp. 4-9

[11] Milam, John H., Jr.,2001, “Knowledge Management for Higher Education”, ERIC Digest ED464520

[12] Nagad, W., Amin, G., 2006, “Higher Education in Sudan and Knowledge Management Applications”, IEEE 0-7803-9521-2/06, pp. 60-65

[13] Nakkiran, N.S., Sewry, D.A., 2002, “A Theoretical Framework for Knowledge Management Implementation”, Proceedings of SAICSIT, pp. 235-245

[14] Natali, A.C., Falbo, R.A., “Knowledge Management in Software Engineering Environments”,available at <http:/www.citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.86>

[15] Nonaka, I., 1998, “Knowledge Creating Company”, Havard Business Review on Knowledge Management, Havard Business School Publishing, Boston

[16] O’Dell, C., Grayson, C.J., 1998, “If Only We Knew What We Know”, Free Press

[17] O’Leary, D.E., “Technologies for Knowledge Assimilation”, available at <

http://www.academic.research.microsoft.com/Publication/6928388/technologies-for-knowledge-assilimation>

[18] Petrides, L.A., 2004, "Knowledge management, Information Systems and Organizations", Educause Center for Applied Research, Vol. 2004, Issue 20

[19] Ranjan, J., Khalil, S., 2007, "Application of Knowledge Management in Management Education: A Conceptual Framework", Journal of Theoretical and Applied Information Technology, pp. 15-25

[20] Schwartz, D.G., Divitini, M. and Brashethvik, T., 2000, "Internet-based Organizational Memory and Knowledge Management", Idea Group Publishing, Hershey, PA.

[21] Sedziuviene, N., Vveinhardt, J., 2009. "The Paradigm of Knowledge Management in Higher Educational Institutions", Inzinerine Ekonomika-Engineering Economics (5), pp. 79-90

[22] Tiwana, A., 2000, The Knowledge Management Toolkit: Practical Techniques for Building a Knowledge Management System, Prentice Hall, New Jersey

[23] Wiig, K.M., 1996, "On the Management of Knowledge", available at http://www.km-forum.org/what_is.htm

[24] Yeh, C.M.Y., 2005, "The Implementation of Knowledge Management System i.9, pp. 35-41

[25] Zack, M.H., 1999,"Managing Codified Knowledge", Sloan Management Review, vol. 40, no. 4, pp.45-59

Mamta Bhusry is a PhD Research Scholar in the field of Knowledge Management in Technical Education She has a B.Tech in Electrical Engineering and M.Tech in Information Technology. She has a total of 19 years of academic and the industrial experience. She is presently working as Associate Professor in the Department of Computer Science engineering at Ajay Kumar Garg Engineering College, Ghaziabad, India. She has published papers on knowledge management and has authored a book on E-Commerce. She has guided various B.Tech and M.Tech level projects.

Dr. Jayanti Ranjan is a PhD from Jamia Millia Islamia Central University, India in the field of data mining and has 16 years of teaching experience. She is presently working as Professor, Information Systems Management at Institute of Management Technology, Ghaziabad, India. She has published various papers on data clustering, data mining, database security, business intelligence, educational technologies that appeared in Emerald Publishers, Inderscience Publishers, World scientific Publishers, Asian Network for Scientific Information and other refereed journals. She is serving on the editorial board for the international journal – Information Technology Journal, Inter disciplinary Journal of Information Knowledge and Management, UK, Journal of Theoretical and Applied Information Technology. She is also the Chairman, International Relations, IMT Ghaziabad.

Performance Enhancement of a Dynamic System Using PID Controller Tuning Formulae

JYOTIPRAKASH PATRA[1], Dr. PARTHA SARATHI KHUNTIA[2]

[1]Associate Professor, Disha Institute of Management and Technology
Raipur, India ,

[2]Professor, Hi-Tech College of Engg. and Technology
Bhubaneswar, India

Abstract

The proportional integral derivative (PID) controller is the most dominant form of automatic controller in industrial use today. With this technique, it is necessary to adjust the controller parameters according to the nature of the process. Thus, for effective control of a HVDC system, for example, specific values need to be chosen for the P, I and D parameters, which will be different for the values required to control, for example, an induction motor drive. This tailoring of controller to process is known as *controller tuning*. Controller tuning is easily and effectively performed using *tuning rules* (i.e. formulae for controller tuning, based on process information). Such tuning rules allow the easy set up of controllers to achieve optimum performance at commissioning. Importantly, they allow ease of re-commissioning if the characteristics of the process change. The paper communicates the results of recent work in the collation of industry-relevant PI and PID controller tuning rules, which may be applied to a variety of applications in power electronics, machines and drives.

Keywords: *PI, PID, Tuning Rules, FOLPD model, IPD model.*

1. Introduction

PI and PID controllers have been at the heart of control engineering practice for seven decades. Historically, the first tuning rule for setting up controller parameters was defined in 1934 for the design of a proportional-derivative (PD) controller for a process exactly modelled by an *integrator plus delay* (IPD) model [3]. Subsequently, tuning rules were defined for PI and PID controllers, assuming the process was exactly modelled by a *first order lag plus delay* (FOLPD) model [4] or a pure delay model [4], [9]. In the wide area covered by power electronics, machines and drives, PI or PID controllers have been considered for the control of DC-DC converters (e.g. [1]), flexible AC transmission systems (e.g. [15]), synchronous machines (e.g.[6]), HVDC systems (e.g. [18]), electric vehicle speed (e.g.[14]) and induction motor servo drives (e.g. [13]). In general, at commissioning, the PID controller is installed and tuned. However, surveys indicating the state of industrial practice report sobering results. For example, in the testing of thousands of control loops, it has been found that 65% of loops operating in automatic mo de produce less variance in manual than in automatic (i.e. the automatic controllers are poorly tuned) [8]. Process performance deteriorates when the controller is poorly tuned; this deterioration may be reflected, for example, in a reduction in energy efficiency and increased environmental emissions. The net effect will be an increase in operating costs and a reduction in overall competitiveness. However, good controller tuning, for example, can allow the recovery of up to 6% of energy costs, in a variety of industries [5]. Thus, there is strong evidence that PI and PID controllers remain poorly understood and, in particular, poorly tuned in many applications. This is surprising, as very many tuning rules exist to allow the specification of the controller parameters. Tuning rules have the advantage of ease of calculation of the controller parameters (when compared to more analytical controller design methods), on the one hand; on the other hand, the use of tuning rules is a good alternative to trial and error tuning. It is clear that the many controller tuning rules proposed in the literature are not having an impact on industrial practice. One reason is that the tuning rules are not very accessible, being scattered throughout the control literature; in addition, the notation used is not unified. It is timely, therefore, to communicate the results of recent work done in the collation of tuning rules, using a unified notation, for continuous-time PI and PID control of single input, single-output (SISO) processes [16], [17]. Such rules may be specified for processes either without or with a time - delay (dead-time) term; such terms arise in voltage source inverters, for example, where a dead-time is required to prevent a shorting condition during switching [12]. Generally, a dead-time term is common; sources of dead-time range from the finite time required for information transmission to application-specific issues, such as the dead time in a motor drive due to imperfect mechanical coupling [13]. Firstly, a brief summary of the range of PI and PID controller structures proposed in the literature, together with the process models used to define the controller tuning rules, is provided. Then, controller architecture and process modeling issues are outlined, followed by the outline of tuning rules for setting up PI and PID controllers, for a number of process models. Finally, conclusions to the paper are drawn. Due to space restrictions, a case study of the application of tuning rules to design a controller for a pilot-scale plant is detailed in the poster presentation accompanying this paper.

2. Controller architecture and process modeling

A practical difficulty with PID control technology is a lack of industrial standards, which has resulted in a wide variety of PID controller architectures. Seven different structures for the PI controller and forty-six different structures for the PID controller have been identified. Controller manufacturers vary in their choice of architecture; controller tuning that works well on one architecture may work poorly on another. Full details are given in [16], [17]; considering the PID controller, common architectures are:

1. The 'ideal' PID controller (Figure 1), given by

$$G_c(s) = K_c\left(1 + \frac{1}{T_i s} + T_d s\right) \qquad (1)$$

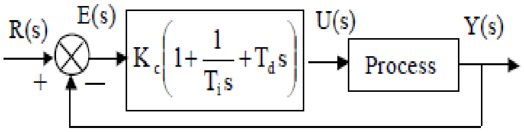


Figure 1. Ideal PID controller in a unity feedback block diagram representation.

This controller structure, and an equivalent structure, is also labeled the parallel, ideal parallel, non-interacting, parallel noninteracting,independent, gain independent or ISA controller [17]. 276 tuning rules have been identified for this controller structure.

This architecture is used, for example, on the Honeywell TDC3000 Process Manager Type A, interactive mode product [11].

2. The 'classical' PID controller (Figure 2), given by

$$G_c(s) = K_c\left(1 + \frac{1}{T_i s}\right)\frac{1+sT_d}{1+s\frac{T_d}{N}} \qquad (2)$$

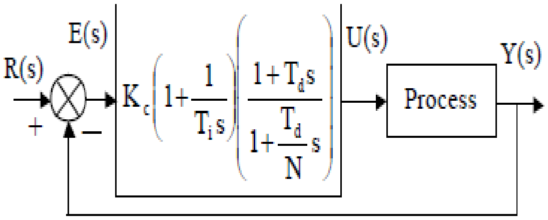

R(s)
E(s)
+
−
$K_c\left(1+\frac{1}{T_i s}\right)\left(\frac{1+T_d s}{1+\frac{T_d}{N}s}\right)$
U(s)
Process
Y(s)

Figure 2. Classical PID controller in a unity feedback block diagram representation. Also labeled the cascade, interacting, series, interactive, rate-before-reset or analog controller [17], 101 tuning rules have been identified for this controller structure.

This architecture is used, for example, on the Honeywell TDC3000 Process Manager Type A, interactive mode product [11].

3. The non-interacting controller based on the two degree of freedom structure (Figure 3), given by

$$U(s) = K_c\left(1 + \frac{1}{T_i s} + \frac{sT_d}{1+s\frac{T_d}{N}}\right)E(s) - K_c\left(\alpha + \frac{\beta T_d s}{1+s\frac{T_d}{N}}\right)R(s) \qquad (3)$$

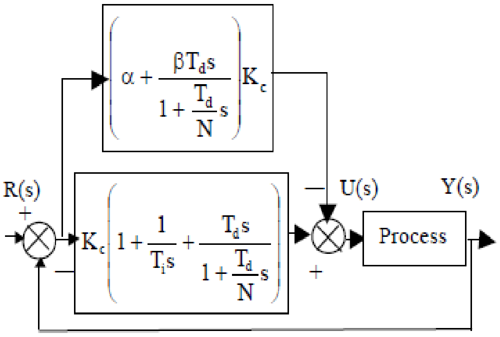

$\left(\alpha+\frac{\beta T_d s}{1+\frac{T_d}{N}s}\right)K_c$
R(s)
+
−
$K_c\left(1+\frac{1}{T_i s}+\frac{T_d s}{1+\frac{T_d}{N}s}\right)$
−
+
U(s)
Process
Y(s)

Figure 3. Non-interacting controller, based on the two degree of freedom structure, in a unity feedback block diagram representation. Also labeled the m-PID or ISA-PID controller [17], 44 tuning rules have been identified for this controller structure.

This architecture is used, for example, on the Omron E5CK digital controller with b =1 and N = 3 [11].

The most dominant PI controller architecture is the 'ideal' PI controller, given by

$$G_c(s) = K_c\left(1 + \frac{1}{T_i s}\right) \qquad (4)$$

The wide variety of controller architectures is mirrored by the wide variety of ways in which processes with time delay may be modeled. Common models are:

1. Stable FOLPD model, given by

$$G_m(s) = \frac{K_m e^{-s\tau_m}}{1+sT_m} \quad (5)$$

2. IPD model, given by

$$G_m(s) = \frac{K_m e^{-s\tau_m}}{s} \quad (6)$$

3. First order lag plus integral plus delay (FOLIPD) model, given by

$$G_m(s) = \frac{K_m e^{-s\tau_m}}{s(1+sT_m)} \quad (7)$$

4. Second order system plus time delay (SOSPD) model, given by

$$G_m(s) = \frac{K_m e^{-s\tau_m}}{T_{m1}^2 s^2 + 2\zeta_m T_{m1} s + 1} \quad (8)$$

$$G_m(s) = \frac{K_m e^{-s\tau_m}}{(1+T_{m1}s)(1+T_{m2}s)} \quad (9)$$

Some 82% of the PI controller tuning rules identified have been defined for the ideal PI controller structure, with 42% of tuning rules based on a FOLPD process model. The range of PID controller variations has lead to a less homogenous situation than for the PI controller; 40% of tuning rules identified have been defined for the ideal PID controller structure, with 37% of PID tuning rules based on a FOLPD process model [17].Of course, the modeling strategy used influences the value of the model parameters, which, in turn, affect the controller values determined from the tuning rules. Forty-one modeling strategies have been detailed to determine the parameters of the FOLPD process model, for example. Space does not permit a full discussion of this issue; further details are provided in [16], [17].

3. Tuning Rules for PI and PID Controllers

Before considering tuning rules for PI and PID controllers in more detail, it is timely to review the action of the PID controller. Consider the ideal PID controller, for example, which is given by

$$G_c(s) = K_c\left(1 + \frac{1}{T_i s} + T_d s\right) \quad (10)$$

With K_c = proportional gain, T_i = integral time constant and T_d = derivative time constant. If $T_i = \infty$ and $T_d = 0$(that is, P control), then the closed loop measured value is always less than the desired value for processes without an integrator term, as a positive error is necessary to keep the measured value constant, and less than the desired value. The introduction of integral action facilitates the achievement of equality between the measured value and the desired value, as a constant error produces an increasing controller output. The introduction of derivative action means that changes in the desired value may be anticipated, and thus an appropriate correction may be added prior to the actual change. Thus, in simplified terms, the PID controller allows contributions from present, past and future controller inputs. PI and PID controller tuning rules may be broadly classified as follows:

- Tuning rules based on a measured step response
- Tuning rules based on minimizing an appropriate performance criterion
- Tuning rules that give a specified closed loop response
- Robust tuning rules, with an explicit robust stability and robust performance criterion built in to the design process
- Tuning rules based on recording appropriate parameters at the ultimate frequency.

Tuning rules in the first four subdivisions are typically based on process model parameters; the development of a process model is typically not required for using tuning rules in the final subdivision above. Some tuning rules could be considered to belong to more than one subdivision, so the subdivisions cannot be considered to be mutually exclusive; nevertheless, they provide a convenient way to classify the rules. An outline of tuning rules in these subdivisions is now provided.

Tuning rules based on a measured step response are also called process reaction curve methods. The first (and most well-known) tuning rule of this type was suggested in 1942 [20]; in this method, the process is modeled by a FOLPD process model with the model parameters estimated using a tangent and point method, as indicated in Figure 4. Simple formulae are used to define tuning parameters for PI and PID controllers. The PI controller settings are given by

$$K_c = \frac{0.9T_m}{K_m\tau_m}, \quad T_i = 3.33\tau_m \quad (11)$$

The (ideal) PID controller settings are given by

$$K_c \epsilon \left[\frac{1.2T_m}{K_m\tau_m}, \frac{2T_m}{K_m\tau_m}\right], T_i = 2\tau_m, T_d = 0.5\tau_m \quad (12)$$

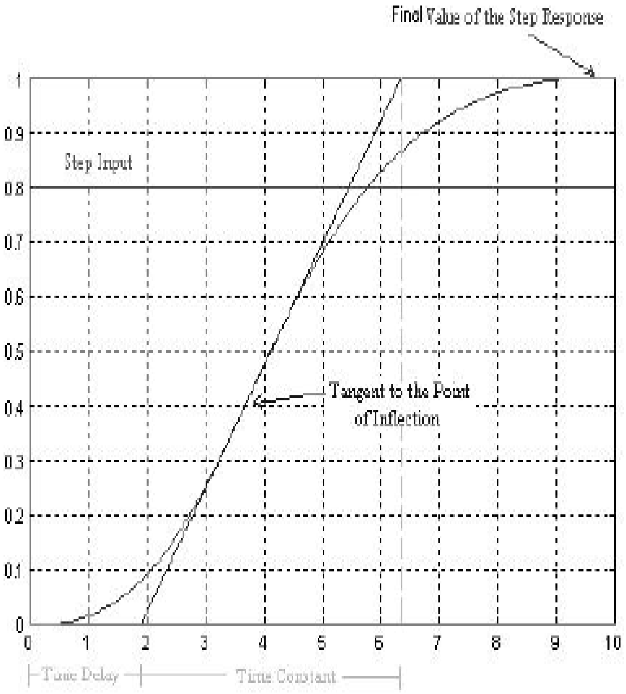


Figure 4. Tangent and point method [20] for developing a process model. Km = model gain = ratio of the steady state change in process output to steady state change in process input, Tm = model time constant and tm =model time delay.

54 controller tuning rules have been identified based on the model parameters determined from this modelling method. 21 of the 47 other modelling methods for determining such a process model, prior to specifying tuning rules, are based on data gathered from the open loop process step or impulse response [17].

Other process reaction curve tuning rules are also described, sometimes in graphical form, to control delayed processes represented by a variety of models [17]. The advantage of process reaction curve tuning strategies is that only a single experimental test is necessary. However, the disadvantages of the strategy are primarily based on the difficulty, in practice, of obtaining an accurate process model; for example, load changes may occur during the test which may distort the test results and a large step input may be necessary to achieve a good signal to noise ratio. Similar disadvantages arise in any tuning method dependent on prior model development.

Tuning rules based on minimizing an appropriate performance criterion may be defined either for optimum regulator or optimum servo action. Performance criteria, such as the minimization of the integral of absolute error (IAE) in a closed loop environment, may be used to determine a unique set of controller parameter values. Tuning rules have been described, sometimes in graphical form, to optimise the regulator response, servo response or other characteristics of a compensated delayed process, represented by a variety of models [17].

Tuning rules that give a specified closed loop response (direct synthesis tuning rules) may be defined by specifying a time domain related metric, such as the desired poles of the closed loop response. The definition may be expanded to cover techniques that allow the achievement of a frequency domain metric, such as a specified gain margin and/or phase margin.

Tuning rules of this type have been specified to compensate a delayed process, represented by a variety of models [17].

Robust tuning rules have an explicit robust stability and/or robust performance criterion built in to the design process. Tuning rules of this type have also been specified to compensate a delayed process, represented by a variety of models [17].

Ultimate cycle tuning rules are based on recording appropriate parameters at the ultimate frequency (that is, the frequency at which marginal stability of the closed loopcontrol system occurs). The first such tuning rule was defined in 1942 [20] for the tuning of P, PI and PID controller parameters of a process that may or may not include a delay. Briefly, the experimental technique is as follows:

a) Place the controller in proportional mode only
b) Increase K_c until the closed loop system output goes marginally stable; record K_c (calling it K_u , the ultimate gain), and the ultimate period, T_u ; a typical marginally stable output, recorded on a laboratory flow process, is shown in Figure 5.

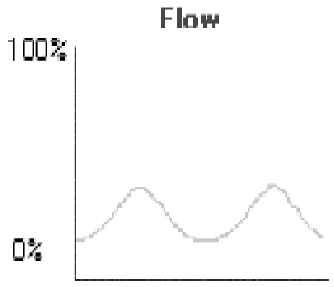


Figure 5. Typical marginally stable process variable pattern. Note that the pattern exhibits evidence of a process nonlinearity, which is common in real applications. Over 129 controller tuning rules have been defined, based on the data determined from such a pattern [17].

Simple formulae are used to define tuning parameters for PI and PID controllers. The PI controller settings are given by

$$K_c = 0.45K_u ,\ T_i = 0.83T_u \qquad (13)$$

with the (ideal) PID controller settings given by

$$K_c = 0.6K_u ,\ T_i = 0.5T_u ,\ T_d = 0.125T_u \qquad (14)$$

The tuning rules implicitly build an adequate frequency domain stability margin into the compensated system [7].
However, there are a number of disadvantages to the ultimate cycle tuning approach:

- The system must generally be destabilized under proportional control
- The empirical nature of the method means that uniform performance is not achieved in general [10]
- Several trials must typically be made to determine the ultimate gain
- The resulting process upsets may be detrimental to product quality
- There is a danger of misinterpreting a limit cycle as representing the stability limit [19] and
- The amplitude of the process variable signal may be so great that the experiment may not be carried out for cost or safety considerations.

Some of these disadvantages are addressed by defining modifications of the rules in which, for example, the proportional gain in the experiment is set up to give a closed loop transient response decay ratio of 0.25, or a phase lag of 135 0 . Ultimate cycle tuning rules, and their modifications, have been specified to compensate general, possibly delayed processes, represented by a variety of models [17].

4. Conclusions

Control academics and practitioners remain interested in the use of PI and PID controllers. PID controller tuning rules can be directly implemented in a variety of applications i.e. the hardware already exists, but it needs to be optimized. The outcome is directly measurable in, for example, energy savings and waste reduction (including greenhouse gas emission reduction). This paper summarizes work carried out in tuning rule development. The most startling statistic to emerge from the work is the quantity of tuning rules identified to date; 443 PI tuning rules and 691 PID tuning rules, a total of 1134 separate rules. Recent years have seen an acceleration in the accumulation of tuning rules. In general, there is a lack of comparative analysis regarding the performance and robustness of closed loop systems compensated with controllers whose parameters are chosen using the tuning rules; associated with this is the lack of benchmark processes, at least until recently [2]. In addition, much work remains to be done in the evaluation of controllers designed using tuning rules in a wide variety of practical applications, including applications in power electronics, machines and drives. The main priority for future research in the area should be a critical analysis of available tuning rules, rather than the proposal of further tuning rules.

Historical note: The 70th anniversary of the receipt of the first technical paper describing tuning rules for setting up controller parameters [4] is presently being marked. The paper was received by the Philosophical Transactions of the Royal Society of London on July 15, 1935; the paper was received, in revised form, on November 26, 1935 and was read on February 2, 1936. The lead author of the paper subsequently took out a patent on the PID controller (Callender, A. and Stevenson, A.B., Automatic control of variable physical characteristics, US patent 2,175,985. Filed: Feb. 17, 1936; Issued Oct. 10, 1939).

References

[1] J. Alvarez-Ramirez, I. Cervantes, G. Espinosa-Perez, P. Maya and A. Morales. “A stable design of PI control for DC-DC converters with a RHS zero”, IEEE Transactions on Circuits and Systems – I: Fundamental Theory and Applications, 46, pp. 103-106, (2000).

[2] K.J. Åström and T. Hägglund. “Benchmark systems for PID control”, Preprints Proc. PID ’00: IFAC Workshop, pp. 181-182, (2000).

[3] A.Callendar. “Preliminary notes on automatic control”, I.C.I. Alkali Ltd., Northwich, U.K., Central File No. R.525/15/3 (1934).

[4] A.Callendar, D.R. Hartree, and A. Porter. “Time-lag in a control system”, Phil. Trans. Royal Society of London Series A, 235, pp. 415-444, (1935/6).

[5] Case histories accompanying Good Practice Guide 346 (Improving the effectiveness of basic closed loop control systems), The Carbon Trust (www.thecarbontrust.co.uk).

[6] H.R. De Azevedo and K.P. Wong. “A fuzzy logic controller for permanent magnet synchronous machine –a sliding mode approach”, Proceedings of the IEEE Power Conversion Conference, pp. 672-677, (1993).

[7] A.M. De Paor. “A fiftieth anniversary celebration of the Ziegler-Nichols PID controller”, Int. J. Elect. Eng.Education, 30, pp. 303-316, (1993).

[8] D.B. Ender. “Process control performance: not at good as you think”, Control Engineering, September, pp.180-190, (1993).

[9] D.R. Hartree, A. Porter, A. Callender and A.B.Stevenson. “Time-lag in a control system – II”, Proc.Royal Society of London, 161(A), pp. 460-476, (1937).

[10] S.-H. Hwang and T.-S. Tseng. “Process identification and control based on dominant pole expansions”, Chem.Eng. Sci., 49, pp. 1973-1983, (1994).

[11] ISMC. RAPID: Robust Advanced PID Control Manual. Intelligent System Modeling and Control nv, Belgium,(1999).

[12] J. Jung and K. Nam. “A PI-type dead-time compensation method for vector-controlled GTO inverters”, IEEE Transactions on Industry Applications, 34, pp. 452-457 (1998).

[13] F.J. Lin, C.M. Liaw, Y.S. Shieh, R.J. Guey and M.S.Hwang. “Robust two-degrees-of-freedom control for induction motor servodrives”, IEE Proc.-Electr. PowerAppl., 142, pp. 79-86, (1995).

[14] S. Matsumura, S. Omatu and H. Higasa. “Improvement of speed control performance using PID type neurocontroller in an electric vehicle system”, Proceedings of the IEEE World Congress on Computational Intelligence, 4, pp. 2649-2654, (1994).

[15] S. Morris, P.K. Dash and K.P. Basu. “A fuzzy variable structure current controller for flexible AC transmission systems”, Proceedings of the IEEE Transmission and Distribution Conference and Exhibition, 1, pp. 330-335,(2002).

[16] A.O’Dwyer. Handbook of PI and PID controller tuning rules. London, U.K.: Imperial College Press, (2003).

[17] A.O’Dwyer. Handbook of PI and PID controller tuning rules (Edition 2). London, U.K.: to be published by Imperial College Press, (2006).

[18] K.R. Padiyar and N. Prabhu. “Modelling, control design and analysis of VSC based HVDC transmission systems”, Proceedings of the International Conference on Power Systems Technology, pp. 774-779, (2004).

[19] D.W. Pessen. “A new look at PID-controller tuning”, Trans. ASME. J. Dyn. Sys., Meas. Control, 116, pp. 553-557, (1994).

[20] J.G. Ziegler and N.B. Nichols. “Optimum settings for automatic controllers”, Trans. ASME, 64, pp. 759-768,(1942).

First Author: I completed my B.E. degree in Computer Science & Engineering from B.P.U.T, Odisha, in the year 2004,then completed M.E.(Computer Technology and Application) from CSVTU, Chhattisgarh ,in the year 2008.I am pursuing my Ph.D. work from MATS University, Raipur. Currently working as a Associate Professor at DIMAT, Raipur. I have already published a book “Analysis and Design of Algorithms” under SUN INDIA PUBLICATION, NEW DELHI.

Second Author: I completed my M.E. (Automatic Control System & Robotics) from M S University, Baroda, in the year 1999 .completed Ph.D. from Indian School of Mines University, Dhanbad, India, in the year 2009.Currently working as Prof. & Head at Hi-Tech College of Engineering and Technology, Bhubaneswar, India. I have already published two books, first one (Recent Advances in Control Systems, Robotics and Automation, International Society of Advanced Research, Third edition, Volume-1, ISBN 978-88-901928-6-9, pp.54-59.) , and second (Development of Intelligent Control Strategies for Aircraft and Other Dynamic Systems “LAP LAMBERT Academic Publishing AG&Co.KG Saarbrücken,Dudweiler Landstrabe 99, 66123 Saarbrucken Germany, ISBN NR-978-8383-8360-6).

Using Bee Colony Optimization to Solve the Task Scheduling Problem in Homogenous Systems

Vahid Arabnejad[1], Ali Moeini[2] and Nasrollah Moghadam[3]

[1] Department of Computer Engineering, Islamic Azad University, South branch
Tehran, Iran

[2] Computer Engineering Dept., University of Tehran
Tehran, Iran,

[3] Computer Engineering Dept., University of Tarbiat Modares
Tehran, Iran

Abstract

Bee colony optimization (BCO) is one of the most recent algorithms in swarm intelligence that can be used in optimization problems this algorithm is based on the intelligent behavior of honey bees in foraging process. In this paper bee colony optimization is applied to solve the task scheduling problem which tasks have dependency with each other. Scheduling of tasks that represents by directed acyclic graph is a NP-complete problem. The main purpose of this problem is obtaining the minimum schedule length that is called make-span. To realize the performance of BCO in this problem, the obtained results are presented and compared with the most successful methods such as Ant colony system, Tabu search and simulate annealing. The comparison shows that BCO produces the solutions in a different way and it is still among the bests.

Keywords: *Bee Colony Optimization, Task Graph, Task Scheduling Problem, Homogenous Processors.*

1. Introduction

One of the most significant, vital, and complex problems of parallel execution is referred as Scheduling a set of either dependent or independent tasks on a set of processors . Parallel programs can be divided into a group of smaller tasks which are usually related to each other. Minimizing of the scheduling length (*make-span*) is known as the only purpose of task scheduling problem in order to allocate tasks to processors such that dependencies between tasks are satisfied.

Task scheduling problem is separated into two groups which are either with or without communication costs, in which each group could be individually proposed in heterogeneous or homogeneous systems.

The algorithms for finding the optimal result for the multiple-processor scheduling problem have been demonstrated to be NP-complete [1, 6].

Many metaheuristic have been proposed based on methods and approaches to the task scheduling [2-5].

Behaviors of Social insects such as ants and bees in the real world have been studied many years to solve many problems. Ant colony algorithm is an example of swarm intelligence algorithms for solving combinatorial optimization problems. Ants can find the shortest path from the food source to their nest by using pheromone [10].

In this paper Task Scheduling Problem has been solved by the bee colony optimization. The bee colony optimization algorithm is inspired by the behavior of a honey bee colony in nectar collection, is another example of swarm intelligence. BCO has been proposed by Lucic and Teodorovic [6-8]. Artificial bees in BCO cooperate to solve combinatorial optimization problem. Every bee during the search process makes some moves and constructs a solution [5]. Furthermore, we add a global memory for bees to compare their result with previous iteration results that will be explained in details later.

2. Definition of Task Scheduling Problem

The problem of task scheduling is indicated by a directed acyclic graph (DAG). This graph is shown by G (V, E, w, c) which has four characters that are:

V is the set of v nodes, and each node $v_i \in V$ represents a task.

W is a V computation costs array in which each w_i gives the estimated time of task execution.

E is the set of communication edges. The directed edge e_{ij} joins nodes v_i and v_j , where node v_i is called the parent node and node v_j is called the child node.

C is the set of communication costs, and edge e_{ij} has a communication cost $c_{ij} \in C$.
The relationship of data-dependency from task t_i to t_j could be indicated via directed edge e_{ij} in the set
E = {e_{ij} | *i*, *j* ∈ {1, 2... |*V* |}. On the other words, task t_i transfers vital relevant information to task t_j after finishing its execution. The amount of data transferred from task t_i to task t_j , is measured by the weight of the edge e_{ij} , which is denoted D(t_i , t_j).
The task t_i is named the *predecessor* for the task t_j , and the task t_j is the *successor* for the task t_i. *Pred*(t_i) denotes a set of its predecessors, and *Succ*(t_i) denotes a set of its successors. In DAG, if a task t_i exists that could satisfy *Pred*(t_i)=φ, it is called the *entry task* and is denoted by t_{entry} . On the other hand, if there is a task t_j that could be able to satisfy the equation of *Succ*(t_j)=φ, this task is called the *exit task* and is denoted by t_{exit} .
Some virtual tasks under the following conditions are added into the DAG, in order to ensure the DAG has only one input and one output tasks. A virtual entry task with zero workload should be joined to the DAG while there are many entry tasks in a DAG. The directed edges from this virtual entry task to each entry task can be established, and the amount of transmission data of these directed edges is zero. On the other words, if there are many exit tasks in a DAG, and then a virtual exit task that has zero workload should be joined to the DAG. The directed edges from each exit task to this virtual exit task are established, and the amount of transmission data of these directed edges is zero, too. Therefore, a DAG that only has one t_{entry} and t_{exit} can be designed [11].
In order to find the finishing time of each node execution, its start time is added with its weight that is [12]:

$$t_f(i) = t_s(i) + w(i) \qquad (1)$$

Two nodes could not be executed on just one processor simultaneously. The costs relationships between the nodes that are executed on a same processor are considered to be zero because these are some local relationships.
The time in which a communication arrives at the destination processor is mentioned as the edge finish time. For a graph G (V, E, w, c) with nodes n_i and n_j and the edge e_{ij} the amount of finishing time for that edge is equivalent with the sum of the completion time of node n_i execution and the weight of the edge e_{ij} [12].

$$t_f(e_{ij}, P_{src}, P_{dst}) = t_f(n_i, P_{src}) + \begin{cases} 0 & if P_{src} = P_{dst} \\ c(e_{ij}) & otherwise \end{cases} \qquad (2)$$

As it could be seen from the equation above, there are two cases for the weight of edge e_{ij} that indicates the relationship's cost: if node n_j is executed on the same processor in which node n_i were processed, or in the other words nodes n_i and n_j have the same processor, the weight of that edge is considered to be zero. Otherwise, the written number on that edge shows its weight and n_j could not be executed as long as n_i that was executed completely. The node n_j will be started to be executed immediately after the completion of node n_i. This problem is known and defined as the problem of priority constraint (limitation). The nearest time that the execution of node n_i could be started is named Data Ready Time and it is indicated by DRT that could be computed through the equation below:

$$t_{dr}(n_j, P) = \max_{n_i \in Pred\, n_j} \{t_f(e_{ij}, proce(n_i), P)\} \qquad (3)$$

If node n_j is an root node $t_{dr}(n_j, P) = 0$
Limitations on the start time of node n could be formulized via DRT:

$$t_s(n, P) \geq t_{dr}(n, P) \qquad (4)$$

A scheduler duty is considered to be completed when the last node of our graph was scheduled and there were not any other nodes for scheduling. If we want to obtain the length of a scheduler it would be:

$$sl(S) = \max_{n \in V}\{t_f(n)\} - \min_{n \in V}\{t_s(n)\} \qquad (5)$$

A target parallel system **P** consists of a set of identical connected processors which has the following properties:

1. All of the processors could execute only a task during its allocated period of time.
2. The amount of communication costs between tasks which are executed on the same processor should be as negligible as the case that it could be presume to zero.
3. The communication network is fully connected, in which every processor could communicate with other processors, directly.

3. **Bee Colony Optimization**

Each bee hive has a place which is called dance floor. Every Bee starts to dance after when it came back to its hive from a foraging. The main purpose of this kind of dancing is to convince the other bees to be accompanied by them. The procedure of finding a food source in the BCO algorithm is separated into 2 steps.
Forward pass: in this step bees leave their hive for finding a proper food source around their hive. A parameter which is called *NC (number of* solution components) is defined here. This parameter determines the number of tasks that must be visited by each bee in its forward pass. Then a partial solution is generated according to the tasks which are visited by each bee in every forward pass procedure.

The amount of NC is determined practically before the process of searching will be started.

Backward pass: all of bees come back to their home in this step and then they start to calculate and evaluate their answers. Afterwards, these answers are compared with each other in order to find the best answer. In other words, each bee should decide to be loyal to its path or not. Each bee could do one the three jobs below when it came back to the hive [13]:

1- It could advertise its own path in order to absorb the other bees.
2- It could leave its path and join to another bee.
3- It could decide not to advertise its own path; however it keeps on its path.

In these step, those bees which generate more appropriate and better results have more chances of success in order to advertise and absorb the other bees. They communicate their obtained information about the quality of partial solution and their results with other bees via their dances. The duration of every bee's dance is highly likely related to the quality of its obtained result.

These two steps are repeated consecutively in order to generate a complete result which is equivalent to execute all tasks in the task scheduling problem. At last, the most proper and best result is chosen.

Below, a pseudo code for the BCO algorithm is written [14]:

```
B: the number of bees involved in the search.
NC: the number of forward (backward) passes in a single iteration.
Do
  1- Initializing
  2- For (i = 0 ; i < NC ; i ++ )
       //forward pass
         a) For (b = 0 ; b < B ; b ++ )
              (1) Evaluate all possible moves;
              (2) Choose one move using the roulette wheel.
       //backward pass
         b) For (b = 0 ; b < B ; b ++ )
              Evaluate (partial/complete) solution for bee b;
         c) For (b = 0 ; b < B ; b ++ )
              Loyalty decision using the roulette wheel for bee b;
         d) For (b = 0 ; b < B ; b ++ )
              If (b is follower), choose a recruiter by the roulette wheel.

  3. Evaluate all solutions and find the best one
While stopping criteria is not satisfied
```

In task scheduling problem, two factors should be considered by each bee in every forward pass:

1) Which task should be selected to execute
2) Which CPU should be chosen to execute the task

First of all, each bee calculates the number of executable tasks, which could be executed when all of their dependent precedence tasks were completed. Then, each bee could peek one of these tasks by consideration of some factors such as duration of task execution, the number of other tasks which are related to a significant task and etc. after choosing a desired task, a proper processor should be chosen by that bee. The probability of choosing a proper processor could be calculated via the formula below:

$$p_j = \frac{1 - T_j}{\sum_{i=0}^{n} T_i} \qquad \text{When j} = 1,2, \dots \text{k} \qquad (6)$$

Where, T_j is the quickest duration of task execution on the jth CPU, and also k is the number of CPUs.

In this algorithm a global memory is defined and considered for all of bees. When each stage of forward pass completes, after the determination of all the bees which have the permission of dance, their results are averaged and saved in the memory. The bees use these saved information after the next iteration wants to be started. While a partial solution is generated, if the average of posterior results is not acceptable in comparison with the prior results, the mentioned way will be forgotten and leaved by bees and they will come back to their hive. The speed of execution would be increased clearly via this method.

After the first step of task scheduling is completed and all the bees come back to their hive, they will start to share their information to the other bees. In this stage the amount of each bee's loyalty to its path could be calculated by the formula below [14, 15]:

$$P_b^{u+1} = e^{-(O_{max} - O_b)/u} \qquad ; \text{b} = 1, 2 \dots \text{B} \qquad (7)$$

Where

u - The forward pass counter (taking values 1, 2… NC)

O_b is calculated by

$$O_b = \frac{C_{max} - C_b}{C_{max} - C_{min}} \qquad (8)$$

C_b is the result of partial solution for the bth bee. Partial solution result means the latest time point of finishing the last task at any processors.

C_{max} and C_{min} are respectively the largest and smallest partial solution results producing by all bees.

3-Result

In this paper the random graph generator is used to test the proposed algorithm. Graphs based on parameters that will describe below, have been produced.

1. N: Number of nodes (tasks) in the DAG
2. Width or Fat: This parameter represents the maximum number of tasks that can be executed simultaneously. It means that the higher value of this parameter will result a higher degree of parallelism.
3. Density: this parameter indicates the numbers of edges between tasks of two levels of the DAG.
4. Regularity: It is the uniformity of the number of tasks in each level;
5. Jump: this factor indicates the maximum number of levels that an edge could go. For example, every edge can connect with other nodes in 4 levels below with jump=4in the DAG.
6. CCR: it is the ratio of the communication cost to computation cost.

These parameters can have different values. We generated about 648 different graphs with combination of these different values that indicate in table [1].

Table 1: PARAMETERS AND THEIR VALUES USED FOR GENERATING DAGS

N	10 - 20 - 30 - 40
Jump	1 - 2 - 4
Width	0.1 - 0.2 - 0.8
Density	0.2 - 0.8
Regularity	0.2 - 0.8
CCR	0.1 - 0.5 - 0.8
Number of processor	4 - 8 - 16

We selected 5 random graphs among all the generated graphs that their details presented in the table below:

Table 2: DAG properties

	With	Density	Regularity	Jump	CCR
DAG1	0.1	0.2	0.2	1	0.5
DAG2	0.1	0.2	0.8	4	0.1
DAG3	0.1	0.8	0.2	4	0.1
DAG4	0.8	0.8	0.8	2	0.8
DAG5	0.2	0.2	0.8	4	0.1

We compared make-span of these selected graphs in our propose algorithm with Ant colony system (ACS), Simulate Annealing (SA), and Tabu Search (TS). Each of these DAGs is executed on 4, 8 and 16 processors.

Table 3: DAG1 result

	Number of task = 30				Number of task = 40			
	TS	SA	ACS	BCO	TS	SA	ACS	BCO
Number of processor =4	497	497	497	497	673	666	666	666
Number of processor =8	442	442	442	442	585	579	579	579
Number of processor =16	493	493	493	493	663	663	663	663

Table 4: DAG2 result

	Number of task = 30				Number of task = 40			
	TS	SA	ACS	BCO	TS	SA	ACS	BCO
Number of processor =4	201	200	201	201	290	290	290	290
Number of processor =8	253	253	258	253	317	317	320	317
Number of processor =16	267	267	269	269	335	335	337	337

Table 5: DAG3 result

	Number of task = 30				Number of task = 40			
	TS	SA	ACS	BCO	TS	SA	ACS	BCO
Number of processor =4	241	241	242	241	288	287	297	292
Number of processor =8	267	267	272	267	332	332	339	332
Number of processor =16	248	250	253	253	285	285	291	291

Table 6: DAG4 result

	Number of task = 30				Number of task = 40			
	TS	SA	ACS	BCO	TS	SA	ACS	BCO
Number of processor =4	165	166	164	165	213	211	212	212
Number of processor =8	143	141	144	143	152	150	153	152
Number of processor =16	98	98	98	98	106	106	108	108

Table 7: DAG5 result

	Number of task = 30				Number of task = 40			
	TS	SA	ACS	BCO	TS	SA	ACS	BCO
Number of processor =4	200	200	207	201	225	225	229	225
Number of processor =8	169	169	170	169	263	263	269	263
Number of processor =16	203	203	209	209	200	200	203	203

Table8: PAIR-WISE COMPARISON OF THE SCHEDULING AGORITHM

		BCO	SA	TS	ACS
BCO	better	*	30%	45%	60%
	equal		55%	35%	30%
	worse		15%	20%	10%
SA	better	15%	*	25%	40%
	equal	55%		60%	45%
	worse	30%		15%	15%
TS	better	20%	15%	*	20%
	equal	35%	60%		65%
	worse	45%	25%		15%
ACS	better	10%	15%	15%	*
	equal	30%	45%	65%	
	worse	60%	40%	20%	

4. Conclusions

Since the Swarm intelligence become one of the interesting methods in Parallel Computing field, a modified version of the BCO algorithm (which is one of the most recent nature inspired algorithms) has been applied for solving task scheduling problem in this paper. It simulates the intelligent behavior of bees when they are faced with a source. The Task scheduling problem is a kind of NP hard problems which cannot be solved with linear algorithms. Thus metaheuristic algorithms become so interesting to employ for solving such problems. BCO has been rarely applied in this field and this application is a new area for it.

There are some novelties in the presented algorithm, and the most important innovation is considering a general memory for all bees, to compare their obtained results with the acceptable results which are obtained previously. Like other metaheuristic methods BCO has demonstrated solid solutions on this problem, and the obtained results has been presented and compared with some other powerful and well known metaheuristic algorithms such as ACS, SA and TS. The BCO solutions are considerably close to SA that is the best scheme, however this mentioned algorithm has better results in comparison with the ACS algorithm. The results of these algorithms are shown and compared in table8. Consequently, BCO could be considered as a suitable solving method in order to face NP hard problems.

References

[1] M. R. Garey, and D. S. Johnson, “Computers and intractability: a guide to the theory of NP-completeness”, W. H. Freeman and Company, 1979.

[2] P. Shroff , “Genetic Simulated Annealing for Scheduling Data-dependent tasks in Heterogeneous Environments” Proceedings of Heterogeneous Computing Workshop, Apr 1996, pp.98-117.

[3] F.A. Omara and M.M. Arafa , “ Genetic algorithms for task scheduling problem” , Journal of Parallel and Distributed Computing 70 (2010) pp 13_22

[4] N. Nissanke, A. Leulseged and S. Chillara, “Probabilistic performance analysis in multiprocessor scheduling”, Journal of Computing and Control Engineering, 2002, Vol. 13, No. 4, pp.171–179.

[5] M . Rapaic, Z. Kanovic and Z. Jelicic, “A theoretical and empirical analysis of convergence related particle swarm optimization “, WSEAS Transactions on Systems and Control, Nov 2009, Vol. 4, Issue 11, pp. 541-550

[6] P. Lucic, D. Teodorovic, “Bee system :modeling combinatorial optimization transportation engineering problems by swarm intelligence”, in preprints of the TRISTAN IV triennial symposium on transportation analysis, Sao Miguel, Azores Islands;2001.

[7] P. Lucic, D. Teodorovic, “Transportation modeling: an artificial life approach”, in: Proceedings of the 14th IEEE international conference on tools with artificial intelligence, Washington,DC;2002.

[8] P. Lucic, D. Teodorovic, “Computing with bees: attacking complex transportation engineering problems”, International Journal on Artificial Intelligence Tools 2003.

[9] E.G. Co_man, Computer and job-shop scheduling theory. In Wiley, 1976.

[10] M. Dorigo, G. Di Caro and L.M. Gambardella, “Ant algorithms for discrete optimization”, Artificial Life, 5:137-172, 1999.

[11] C.chaing , Y.Lee , C.Lee and T.chou , “Ant colony optimization for task matching and scheduling” , Computers and Digital Techniques, IEE Proceedings - Nov. 2006 Volume: 153 Issue: 6.

[12] O. Sinnen , Task Scheduling for parallel system , Wiley,2007.

[13] D.Teodorovic, M. Dell’Orco, “Bee colony optimization: a cooperative learning approach to complex transportation problems”, in Advanced OR and Al. Methods in Transportation, 2005, pp. 51-60.

[14] T.Davidovic , D.Ramljak, M.Selmic and D.Teodorovic , “Bee colony optimization for the p-center problem” , Computers & Operations Research 38 (2011) 1367–1376.

[15] T.Davidovic, M.Selmic, D.Teodorovic, “Scheduling Independent Tasks: Bee Colony Optimization Approach”, 17th Mediterranean Conference on Control & Automation Makedonia Palace, Thessaloniki, Greece June 24 - 26, 2009.

.

Enhancing Information Systems Security in Educational Organizations in KSA through proposing security model

Hussain A.H. Awad and Fadi M. Battah

Department of Computer and Information Technology, King Abdulaziz University, Faculty of Science and Arts- Khulais Jeddah, KSA

Department of Computer and Information Technology, King Abdulaziz University, Faculty of Science and Arts- Khulais Jeddah, KSA

Abstract

It is well known that technology utilization is not restricted for one sector than the other anymore, Educational organizations share many parts of their information systems with commercial organizations.

In this paper we will try to identify the main characteristics of information systems in educational organizations, then we will propose a model of two parts to enhance the information systems security, the first part of the model will handle the policy and laws of the information system, the second part will provide a technical approach on how to audit and subsequently maintain the security of information system.

Keywords: *Information Systems, Security, Model, Enhancing Security, Security Policy.*

1. Introduction

According to Encyclopedia Britannica; a university is an institution of higher education and research, which grants academic degrees in a variety of subjects. A university is a corporation that provides both undergraduate education and postgraduate education. The word university is derived from the Latin universitas magistrorum et scholarium, roughly meaning "community of teachers and scholars."

With the higher competitive environment around the world, educational organizations are not saving any effort to provide the best educational experience. This effort includes employment of latest technologies available; from the entrance of computers mid 20th century up to the outsourcing of complex Enterprise Resource Planning systems and usage of cloud computing. This usage along side with the expanded branches of educational organizations have presented new challenges, the virtual private networks, wide area networks and usage of web interfaces all together made the educational organizations target as same as any other organization on the cyber space.

According to WhiteHat website security statistics report 2011 ‘’*Most websites were exposed to at least one serious vulnerability every day of 2010, or nearly so (9–12 months of the year). Only 16% of websites were vulnerable less than 30 days of the year overall.”* And *“71% of Education, 58% of Social Networking, and 51% of Retail websites were exposed to a serious vulnerability every day of 2010”*.

In this paper we will tackle the issue of Information Systems safety in educational organization, considering King Abdulaziz University as a case study and propose a two tier model for enhancing the security of information system in educational organizations..

2. Information Systems Security

‘’Information system security relates to the adequacy of management controls to prevent, avoid, detect and recover from whole range of threats that could cause damage or disruption to computer systems.’’ (Pattinson, 2008), the process of information security cannot provide a complete prevention, avoidance, detection and recovery from the threats over it (Singh, 2008). And any self aware Information Systems Management realize that; but the fact that any action of information security management can help to reduce these factors gives that motive to embrace all strategies, models and techniques to achieve that.

We can identify the main process of the information system security in the following diagram based on the previous definition:

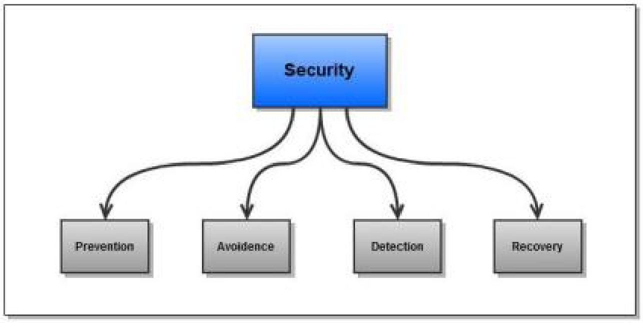


Figure 1 The Main Process of IS Security

2. Universities and Information Systems

“A university information system has to provide information about research and scientific cooperation offers, education and further education capabilities.”(Kudrass, 2006). Information systems in universities can be considered more complex than the usual information systems used in commercial organization. But still it must pay the same attention to its customers (students and members) (Luo and Warkentin, 2004).

2.1 King Abdulaziz University System

The complexity In the King Abdulaziz University information system is relative to The Land Grant University System (Chae and Poole, 2009) and comprised of the following main components:

A- The Students Systems that include
   1- On Demand University Services (ODUS)
   2- Electronic Report System (ERS)
   3- Virtual Classes System (CENTRA)
   4- Electronic Management of Education System (EMES)

B- The Academic Systems that include
   1- On Demand University Services (ODUS)
   2- Academic Affairs System (SMART)
   3- Academic Services for Higher Education
   4- Anjez system for human resources, financial management and memo’s.
   5- Performance Management System (PMS)
   6- Evaluation System

C- Management Systems that include:
   1- Anjez system
   2- Employment system
   3- Decisions and memos system.
   4- Performance Management System (PMS)

The above systems can be viewed by each member (Student, Academic or Employee) depending on his unique number and password. Those information systems are supported by the infrastructure used in the university to provide connectivity amongst campuses in the kingdom and provide internet services for the users.

2.2 Risk Analysis

The main subject of this analysis is to identify to what extent information is subject to change or exposure in this system (Pattinson, 2008), this identification poses the questions of what the possibility of such threats? And what are the expected losses upon such change or exposure?
In real situation it hard to answer these questions, according to (Janczewski, 2009) the reasons behind that are:

- The cost and duration to collect such probabilities may be so huge that job will not be acceptable to management.
- Attacks never happened, but they may happen in the future, so there is no reliable loss of information.

Another factor that raises the risk is the nature of educational organizations, since university standings could suffer greatly from any sort of data manipulation or exposure.
We can provide theoretical risk assessment depending on the nature of the system; the system in the university depends mainly on unique user names and password and on data acquisition upon transfer, especially among long distance branches.
Further investigation to risk analysis requires elaborate software applications which are not our purpose for this paper.

3. Information Systems Security Model

In this section we will tackle the main issue of our paper. Although methodologies are yet till now still not completely mature (Torres, 2009) and some models were proposed to for enterprise security systems (Mazumdar, 2009); we will propose a two part model that would help in the enhancement of information security in universities. This model is divided into two parts; this first is the Policy part, and the second is the Auditing part.

3.1 Policy

We will use the policy based management (Perez, 2006) in order to create an Information System Security (ISS) policy will be aimed to provide the supporting background to regulate the usage by members of the system in a way that raises the system safety and information rigidity to aim in the whole purpose of accomplishing the university strategic goals.

The ISS Policy will be made of the following main points:

1- Statement of Purpose: the statement will in coherence with flexibility, political simplicity and criterion orientation (Baskerville and Siponen, 2002), (Ghormley, 2009).

2- Policy Application Plan: since we are applying new policy to an existing system; a plan must be made for the process of applying this policy, this plan must be governed by time schedule. And it must be made clear that all members comply with this policy (Hinson, 2009).

3- Policy and Standards

1- Overview

2-Responsibility Delegation: Identify the responsibilities for all members by providing security agreement upon using the organizations system. And identify the main manager of the security by appointing Chief Security Officer (CSO) and enable him to form his department accordingly.

3-Contingency Policy: provide plans and walkthrough in case of information disaster, this include performing risk analysis and assessment, business impact analysis, prepare and apply backup and recovery strategy and maintain information update in the policy.

4- Copyrights policy: all members must be informed and agree on the copyright policy for the resources provided in the educational organization system, these copyrights abide to local and international laws. Also they are in line with information technology politics (Petrides, Khanuja-Dhall, & Reguerin, 2006)

5- Help Disk: provide a help disk and hotline to help members in any case of data loss either physical or digital, malfunction of tools, malware and viruses and cases of physical robbery to equipment that holds sensitive data.

6- Accounts and Passwords: members must agree to protect their accounts and not to share any personal information with other members or outside individuals; also the organization must enforce high security policy for selecting personal account passwords, such as using special characters and password change periods.

7- Equipment and Facilities Security: the organization must provide all required equipment, personnel and facilities to protect the main hardware of the system. Also any part of the system design can be count as equipment (Janczewski, 2009).

8- Networking: the organization must provide all necessary tools and equipment to deliver communication network to all university facilities, in the case of KAU this includes colleges outside the city; which means using the latest technology in fiber optics and wireless communication in order to keep all members connected through the main network.

9- Web Security: the organization must work on employing the best security for the webpage since it is the interface used outside the university network and used to access mostly all data.

10- Personal Computers and Laptops: the organization must govern the connectivity to the network using several technologies such as active directory and router identification of computer by MAC address, this will help keep the network secure in case of any attempt to connect unauthorized computer to it.

11- Data Backup: the organization well provides a scheduled process for backup, alongside with providing data containers for the members to store their information on the university equipment as redundant copy.

12- Inventory and turn over policy: the department must write down policy for keeping track of all equipment and the method of upgrades, maintenance and replacement in case of damage, this includes advanced methods of data disposal of old data storage.

13- Users Ethics: a policy of ethical usage of the resources and data provided through the university network must be prepared and signed by all members of the organization. This will help in raise the value of IS in general (Kizza, 2008).

This policy can be furthered in relation with ISO 2700 (Calder, 2006) for information security management system and must be checked regularly in order to make sure it is up-to-date with the latest standards and technologies (Tong, 2009)

3.1 Auditing

This is the second part of our model, this part is concerned of the actual daily vulnerabilities that could happen, in order to do that an audit of our system must be done

periodically, the less the period the better; but due to the nature of our organizations it is hard to conduct such tests and audits in a way that could compromise the system.
In this part we propose a schema for the auditing process, together with the policy this will provide a solid infrastructure for information system security in educational organizations, this process is aimed for the internal use of the information systems department under the approval of higher management only.

3.1.1 The Plan

Any process needs planning, and so in our auditing phase, where this plan will help in the total overview of our process and evaluation of it at the end of each audit process.
A: Establish Goals.
We must at first determine the goals of this process, mainly is to find and secure any vulnerability in our system, these could happen due to: human, software or hardware factors, all must be considered. This plan also must identify clearly the time schedule of this process in order to maintain daily processes ongoing and not interrupting them.
B: Identify Targeted System.
Information about the targeted system is very important in order to identify the size of threats and vulnerabilities, networking protocols, networking schemes, operating systems, management system used and even the forum management software all must be identified.
C: Create Audit Standards. International standards and regulations could be taken in consideration (D'Arcy and Hovav, 2009).
D: Select Security Assessment Tools.

3.1.2 Methodology

A: Preparation.
This is the data collection stage, in which the auditing team works on any information available about the organization.
B: Scanning the system.
Starting with the default ports used then expanding to least used ports; this will help on identifying the entrance to the system.
C: Classify vulnerabilities
D: Break in the system.
Using all information gathered; the team will enforce penetration to the system in order to start the next phase of acquiring information.

3.1.3 Using Social Engineering to Acquire Data

Other than the machines and software; humans are main component in any system (Kuusisto, 2009), and such compenet need some research (Ada, 2009).After using information and wholes of the system; the next step is acquiring data from its containers, this needs more effort using social engineering in order to override: Physical security, Passwords, Network Security, Data Base and Storage Systems

3.1.4 Curtain Down

In this stage all information about the system is analyzed, all holes and vulnerabilities are acknowledged in order to create a plan to plug all the security holes of the system.

4. Summary

In this paper we have proposed a model that can be used to enhance information system security in educational organizations, this model is divided into two main parts, the first part is the policy making and publishing; this part is done thoroughly for the first time then a review and enhancement is done. The second part is the security auditing process; this part must be performed periodically keeping in mind the updates in the software used and any new parts to the system.

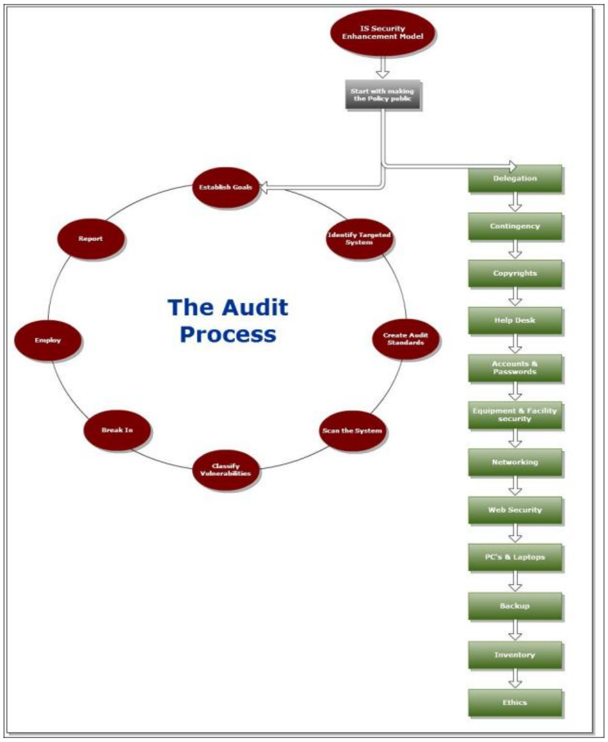


Figure 2 Information Security Enhancement in Educational Organizations Model

References

[1] Abou Bakar Nauman, Romana Aziz, and A.F.M. Ishaq (2009). Selected Readings on Strategic Information Systems (pp. 251-275)

[2] Bongsug Chae, and Marshall Scott Poole (2006). Cases on Information Technology: Lessons Learned, Volume 7 (pp. 388-406)

[3] Carrison K.S. Tong, and Eric T.T. Wong (2009). Governance of Picture Archiving and Communications Systems: Data Security and Quality Management of Filmless Radiology (pp. 53-70)

[4] Chandan Mazumdar (2009). Handbook of Research on Social and Organizational Liabilities in Information Security (pp. 118-132)

[5] Craig Van Slyke (2008). Information Communication Technologies: Concepts, Methodologies, Tools, and Applications. Information Science Reference

[6] Gary Hinson (2009). Handbook of Research on Social and Organizational Liabilities in Information Security (pp. 307-324)

[7] Gregorio M. Perez, Félix J.G. Clemente, and Antonio F.G. Skarmeta (2006). Web and Information Security (pp. 173-195)

[8] Hamid Nemati. ‘’ Information Security and Ethics: Concepts, Methodologies, Tools, and Applications’’, the University of North Carolina at Greensboro, USA. 2008

[9] John D’Arcy, and Anat Hovav (2009).’’ Handbook of Research on Information Security and Assurance ‘’(pp. 55-67)

[10]Jose M. Torres (2009). Handbook of Research on Information Security and Assurance (pp. 467-482)

[11]Joseph Kizza, and Florence Migga Kizza (2008). Securing the Information Infrastructure (pp. 336-354)

[12]Juha Kettunen (2009). Encyclopedia of Information Communication Technology (pp. 542-547)

[13]Lech Janczewski (2009). Encyclopedia of Multimedia Technology and Networking, Second Edition (pp. 1249-1256)

[14]Lisa Petrides, Sharon Khanuja-Dhall, and Pablo Reguerin (2006). Cases on Information Technology and Organizational Politics & Culture (pp. 45-55)

[15]Rauno Kuusisto, and Tuija Kuusisto (2009). Social and Human Elements of Information Security: Emerging Trends and Countermeasures (pp. 77-97)

[16]Serkan Ada (2009). Handbook of Research on Social and Organizational Liabilities in Information Security (pp. 279-292)

[17]Xin Luo and Merrill Warkentin, ‘’Assessment of Information Security spending and costs of failure’’, Mississippi State University, 2004.

[18]Yvette Ghormley (2009). Handbook of Research on Information Security and Assurance (pp. 320-330)

Hussain A.H Awad received his B.Sc. in Business Administration from Mutah University in Jordan, and obtained his M.Sc in Management Information Systems from Amman Arab University in Jordan, and Ph.D. in Management Information Systems from the Arab Academy for Banking and Financial Sciences in Jordan. He is now an Assistant Professor at the Department of Management Information Systems in KING ABDULAZIZ UNIVERSITY, Faculty of Science and Arts - Khulais, Jeddah - Kingdom of Saudi Arabia. His current research interests are supply chain management, IS security, and information retrieval.

Fadi M. Battah Holds a BSc. In Computer Science and M.Sc. in Information Technology Management from University of Sunderland, UK. And has worked as an IT Manager in private sector and now working as a Lecturer at the Department of Management Information Systems in KING ABDULAZIZ UNIVERSITY, Faculty of Science and Arts - Khulais, Jeddah - Kingdom of Saudi Arabia. His current research interests are Information Security, Networking Management, Outsourcing and Information Technology Management.

Process of Reverse Engineering of Enterprise Information System Architecture

Mohammed Abdul Bari[1], **Dr. Shahanawaj Ahamad**[2]

[1 & 2] **Department of Computer Science, College of Science & Arts**
University of Al-Kharj
Wadi Al-Dawasir-11991
Kingdom of Saudi Arabia

Abstract

The availability of computer-aided systems-engineering environments has redefined the organization system development. Pace of change accelerates in the twenty-first century as a result of technological opportunities, liberalization of the world markets, demands for innovation and continually decreasing life cycle of software. Organizations have to continuously re-adjust and re-align their operation to meet all these challenges. This pace of changes has increasingly forced organization to be more outward looking, market-oriented and knowledge driven. This paper present integrated framework of how reverse engineering process work goes in Enterprise Information System(EIS).The essential functions in reverse engineering, how they are associated with EIS, what will be the impact on the organization which is using reverse engineering.

1. Introduction

Is a professional discipline that bridges the business field and computer science field which evolve towards a new scientific area of study. [1, 2, 3, 4]. Silver et al [5] define information system are implemented within an organization for the purpose of improving the effectiveness and efficiency of the organization. The term information often refers as the interaction between algorithmic process and technology. The interaction can occur within or across organizational boundaries. Information system is not only the technology an organization uses, but also the way organization interact with technology and the way the technology works with organization business process. Information system is distinct from information technology (IT), it the component used by information system that interacts with process components. Typically, Information systems include people, procedures, data, software and hardware that are used to gather and analyze digital information. [6, 7]. Overall, an information system discipline emphasized functionality over design. [8]. Over the last 5-8 years numbers of artifacts exploring manipulate, analyze, summarize, hyperlink, synthesize and componentized of the reverse engineering. Many reverse engineering tools on getting on the knowledge from the old software and transfer this information into the minds of software engineers to reuse it.

2. Enterprise Information System (EIS)

Generally kind of computing system dealing with large volumes of data and is capable of supporting information of large organization [9]. EIS provide a technology platform that enables organization to integrate and co-ordinate their business process. It provides a single central system to the organization and ensures that information is shared across all functional level. A typical enterprise information system would be housed in one or more data centers, which runs enterprise software, that include application software that runs across organization borders, such as content management system.

The term “enterprise” is used in many circumstances [10]

- An entire business process.
- A part of large organization.
- A multiple outsourced business operation

It also include the social-technical system [11]

- People.
- Information.
- Technology.
- Business.

2.1 Enterprise Architecture (EA)

Enterprise architecture uses various business method, analytical techniques and conceptual tools to understand the dynamics of an enterprise in doing so, they produce lists, drawing, documents models which together called artifacts [12]. The artifacts describe organization business functions, business capabilities, business process, information resource, software application, and information exchange and communication infrastructures within the enterprise. The collection of artifacts completes enterprise architecture [13].The enterprise has been shown in figure 1.

The Enterprise Architecture is divided into 4 domains as shown in figure 1.

Domain 1:

- Business Architecture:
  - Business strategies are taken , goals are maps with the help of operation model [14,15]
  - Functional decompositions
  - Business process, workflow, rules that assign authorities [16].
  - Organization cycle.

Domain 2:

- Application Architecture :
  - This domain mainly acts interface between applications taking like messages, events.

Domain 3:

- Information Architecture :
  - Data architecture: It describes how the data is processed, how it is stored, how the project team used the data.
  - Master data management: It ensures that an organization does not use multiple versions of the same data in different parts of its operation, which is common in large organization.
  - Metadata: It describes enterprise data elements.

Domain 4:

- Technology Architecture :
  - It deals with application execution environments, operating framework which include application server environment ,operating system, authentication and its environment , security systems and other things,
  - Hardware platform, local and wide area network connection, internet connectivity diagram and programming languages.

3. Reverse Engineering

It is a process of analyzing required system's components, their interrelationships to create representation of the system in another form [17]. Reverse engineering often involves an existing system as its subject, where we can perform reverse engineering starting from any level or at any stage of the life cycle. It covers a broad range starting from the existing implementation, recreating the design, deciphering the requirement actually implemented by subject system. The figure 2 is shown, how reverse engineering process work in Enterprise.

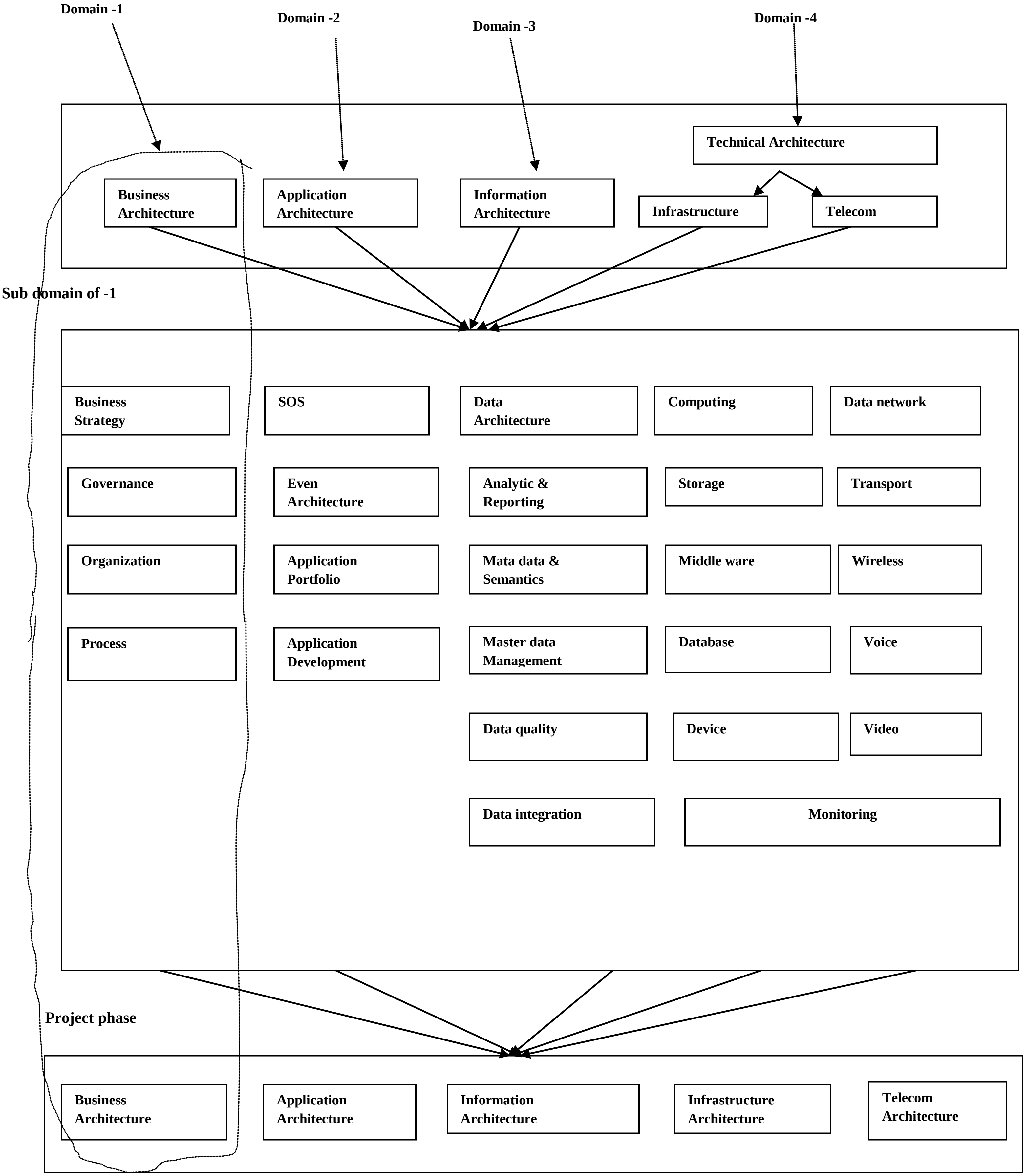


Fig 1: Enterprise Architecture [10]

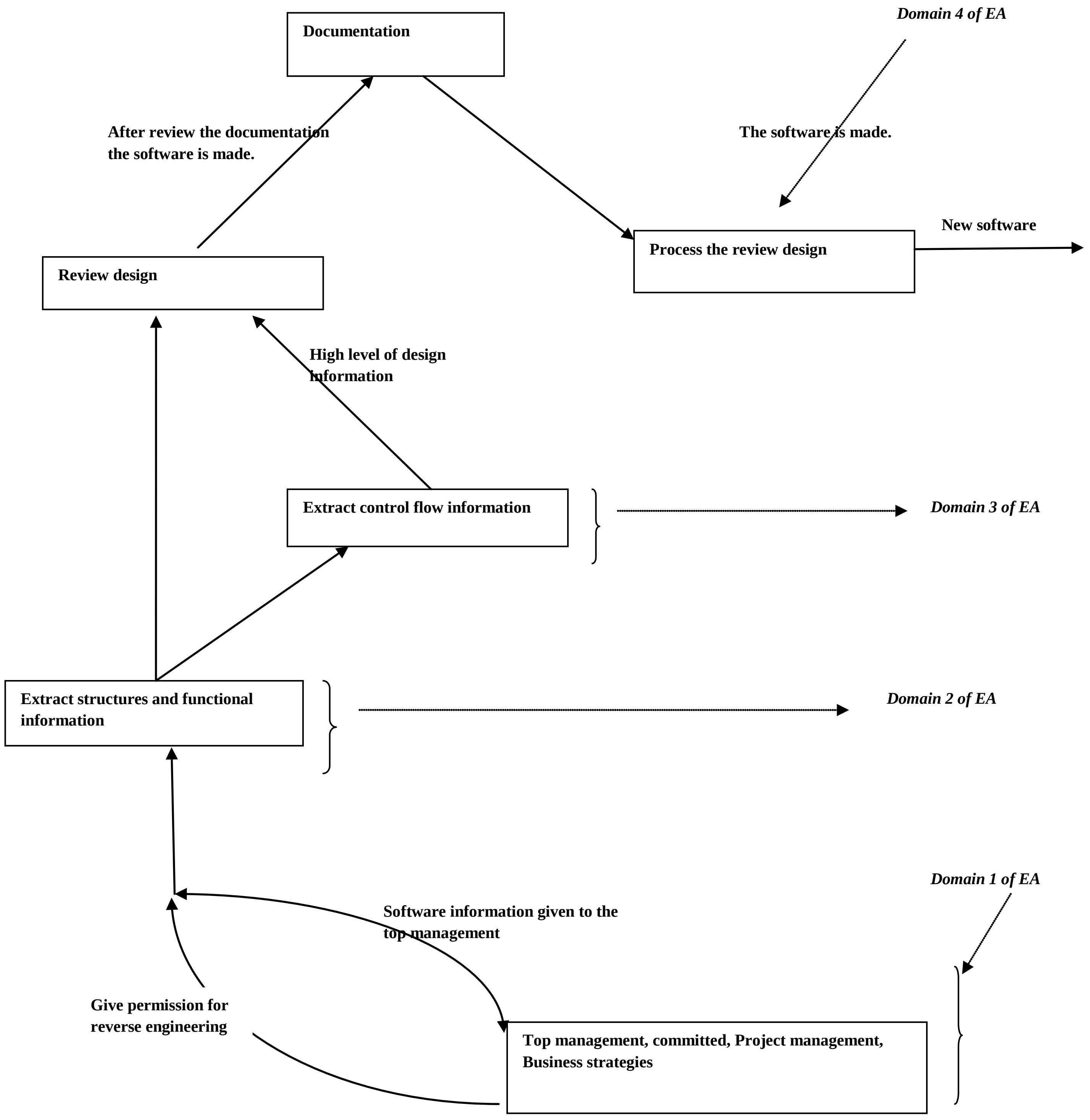

Documentation
Domain 4 of EA
After review the documentation the software is made.
The software is made.
New software
Process the review design
Review design
High level of design information
Extract control flow information
Domain 3 of EA
Extract structures and functional information
Domain 2 of EA
Domain 1 of EA
Software information given to the top management
Give permission for reverse engineering
Top management, committed, Project management, Business strategies

Fig 2: Reverse engineering process in Enterprise

- Top management, business requirement, business strategies make the organization to go ahead with reverse engineering process, which the organization is committed to their share holders and customers.
- All possible information about the software including source code, documentation for

the system cell, personnel experienced (required person) should be identify. This steps allows the software engineer(s) doing the recovery become familiar with the system and its components. This is also the domain -2 of the EA which explain the way the data will be processed.

- Identifying the structure of the software and used to this to create a chart where each node represent function of the software, record the processing done by each node on a PDL [18] can be used to represent functionality routine, analyzed the PDL to identify data transformations in the software .Then identify the high –level control structure of the software which can be record by using control-flow diagram [19]. The high –level control will give an idea of overall operation for the software , which is nothing but the domain-3 of the EA ,where all inventories ,diagram, functions line are represented .
- In review see all the information of the software is available, identify any missing functions, and review the design again.
- After reviewing, the next step is to generate design documentation .Information explaining the business advantage of the software, software overview, its history and the complete software design steps.

- The final step is to process the software, which is nothing but make the software based on the documentation provided to the software engineer(s). These also include application execution environment, operating system, platform software and other things which is nothing but the domain - 4 of the EA.

4. Impact on organization

Today especially large organizations are not only faced with the problem of replacing their information system with the new ones, but they have to maintain again. Reverse engineering provide the means for this purpose supporting in recapturing lost information, restructuring complex system to new one and more maintaining the architecture .The pressure from the market and from the stakeholder’s has let to extremely short product life cycle, most of the released software product never comes in the maintenance phase but instead of being maintained, they are replaced by the new one. For this, too much emphasis is been given on the early phases of the software life cycle, covering only the development part of the software. Large organization, particularly those in fast growing high-tech companies are looking for alternative strategies to minimize their risks and also product new software product, for this companies are seriously think of reverser engineering as the technology [20,21]. Reverse engineering not only increase the structure. It will break down the hierarchical structures, freeing the people to be more innovative and more flexible.

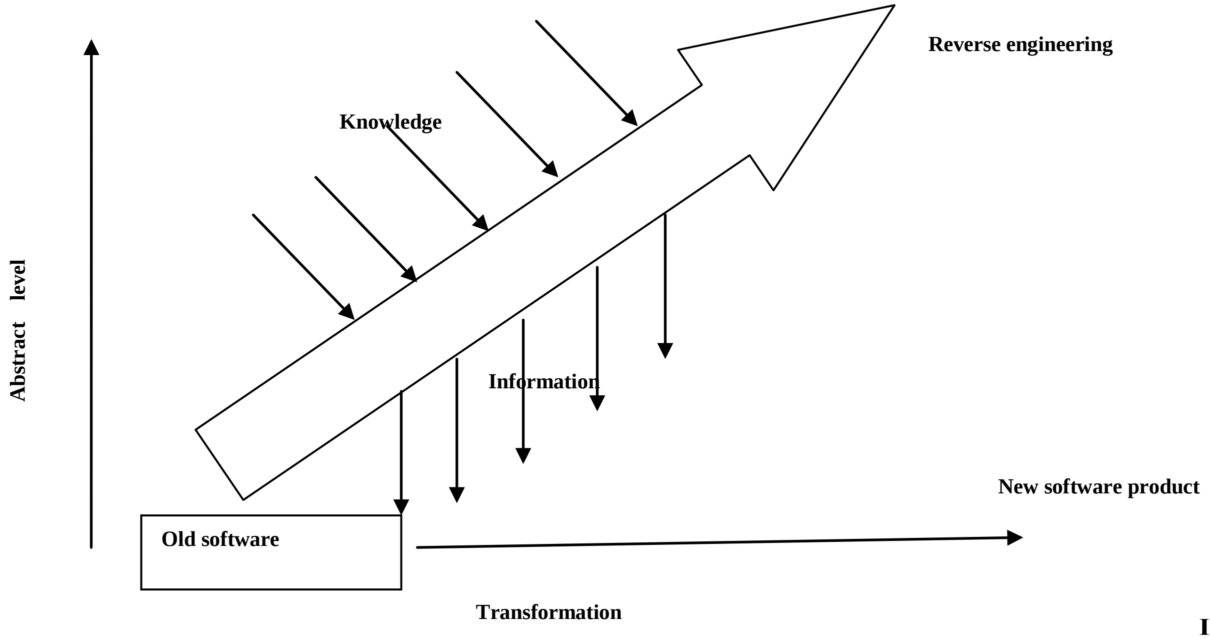


Fig 3: Reverse engineering used in organization

5. Conclusion

The field of IT support system has moved away from stand-alone, dedicated solution with localized impact to more integrated, flexible enterprise-wide system, a fresh approach was needed. In essence enterprise architecture bring with it. Not only it addresses organizational business change perspective but also support the software configuration. This paper has provided a frame work for examining reverse engineering process with the help of Enterprise architecture. Reverse engineering is rapidly becoming a recognized and important component of the future software environments. It provides a major link in the overall process of software development and maintenance. Reverse engineering, used with evolving software development will provide significant enhancement to new software product.

The proposed architecture of reverse engineering (which has taken enterprise architecture in to account also) human expertise is used on well defined occasion to overcome statement situation because of hybrid nature and goal oriented approach. The most promising direction in this area is the continuous software understanding approach. The premise the software reverse engineering need to applied continuously throughout the lifetime of the software and it is important to understand the potential reconstruct the early design and the architectural decision. For the future it is critical that we can effectively answer questions such as “how much knowledge, what level of abstraction do we need to extract from the software, to make informed decision about reengineering”. It will be never able to predict all needs of the reverse engineering and their fore, must developed tools are the end user programmable. Software architecture is now established in many computer science course but the topics such as software evolution, reverse engineering and software migration are rare.

References

[1] Archibald, J.A. "Computer science education for majors of other disciplines". (May 1975). AFIPS Joint computer conferences: 903–906.

[2]Denning, Peter, "Computer science the discipline”. Encyclopedia of Computer Science (2000 Edition).

[3]Coy, Wolfgang. "Between the disciplines". (June 2004), *ACM SIGCSE* Bulletin 36 (2): 7–10.

[4]Hoganson, Ken."Alternative curriculum models for integrating computer science and information ".(December 2001) Journal of Computing Sciences in Colleges 17 (2): 313–325.

[5]Mark S. Silver, M. Lynne Markus, Cynthia Mathis Beath ,” The information technology interaction model: A foundation for the MBA core course”, (Sep,1995) *MIS Quarterly*, Vol. 19, No. 3

[6]Wikipedia ,” Information architecture “ ,(2010), http://en.wikipedia.org/wiki/Information_architecture

[7]Wikipedia,” Information system”,(2010), http://en.wikipedia.org/wiki/Information_systems

[8]Freeman, Peter; Hart, David, "A science of design for software-intensive systems computer science and engineering", (2004), Communications of the ACM 47 (8): 19–21

[9]Wikipedia, ”Enterprise information system”,(2010), http://en.wikipedia.org/wiki/Enterprise_Information_System

[10]Wikipedia ,”Enterprise architecture,” (2010) http://en.wikipedia.org/wiki/Enterprise_architecture

[11]Giachetti, R.E., “Design of enterprise systems, theory, architecture, and methods”, CRC Press, Boca Raton, FL, (2010).

[12]Pelle Ehn ,”Work-oriented design of computer artifacts”. Erlbaum Associates Inc. Hillsdale, NJ, USA,(1990)

[13]Spewak, Steven H. and Hill, Steven C. “Enterprise architecture planning - Developing a blueprint for data applications and technology”, John Wiley,(1992).

[14]Richard Lynch, John Diezemann and James Dowling, "The capable company: Building the capabilities that make strategy work", Wiley-Blackwell,(2003)

15]Bruce R. Scott, "Stages of corporate development (Part I)" ,Harvard Business School Note 371-294.

[16]Howard Smith and Peter Fingar," Business process management", The Third Wave, MK Press(2003).

[17]Elliot J.Chikofsky ,James H.Cross N," Reverse engineering and design recovery : A Taxonomy ,",IEEE,(1990).

[18]Stephen H.Caine ,E.Kent Gordon ,"PDL: a tool for software design", AFIPS '75 Proceedings, (1975).

[19]Ward.P.T,"The transformation schema :Data flow diagram to represent control and timing ",IEEE,(1986)

[20]H.M.Sneed,," Economics of software engineering ", journal of software maintenance , (Sep 1991)

[21]H.M.Sneed and A.Kaposi," A study on the effect of re-engineering upon software maintainability", IEEE Computer society press, (1990).

[22]Hausi A.Muller, Jens H.Jahnke, Dennis B.Smith ,Margarat-Anne Storey , Scott R.Tilly,Kenny Wong,"Reverse engineering : A road map ",

[23]Mohammed Abdul bari, Dr.Shahanawaj Ahmed,"Managing knowledge in development of agile software ",IJACSA,(2011)

About the Authors:

Mr. Mohammed Abdul Bari is an Information System Architect and expert in handling software process improvement. His research area includes Business Process Reengineering, Process Modeling, Information System Redesign and Reengineering. He did B.E. in Computer Science & Engineering from Bangalore University, INDIA and M.S. in Information Systems from London South Bank University, United Kingdom, currently pursuing Ph.D. in Computer Science from University of Newcastle, District Columbia, U.S.A.

Dr. Shahanawaj Ahamad is an active academician and researcher in the field of Software Reverse Engineering with experience of ten years, working with Al-Kharj University's College of Science & Arts in Wadi Al-Dawasir, K.S.A. He is the member of various national and international academic and research groups, member of journal editorial board and reviewer. He is currently working on Legacy Systems Migration, Evolution and Reverse Engineering, published more than twenty papers in his credit in national and international journals and conference proceedings. He holds M. Tech. followed by Ph.D. in Computer Science major Software Engineering, supervised many bachelor projects and master thesis, currently supervisor of Ph.D. theses.

Reliable Communication in Wireless Body Area Sensor Network for Health Monitoring

Saeid Bahanfar[1], Ladan Darougaran [2] , Helia Kousha[3] and Shahram Babaie[4]

[1] Department of Computer Engineering, Tabriz Branch, Islamic Azad University, Tabriz, Iran

[2] Department of Computer Engineering, Tabriz Branch, Islamic Azad University, Tabriz, Iran

[3] Department of Computer Engineering, Tabriz Branch, Islamic Azad University, Tabriz, Iran

[4] Department of Computer Engineering, Tabriz Branch, Islamic Azad University, Tabriz, Iran

Abstract

Now days , interests in the application of Wireless Body Area Network (WBAN) have grown considerably. A number of tiny wireless sensors, strategically placed on the human body, create a wireless body area network that can monitor various vital signs, providing real-time feedback to the user and medical personnel. This communication needs to be energy efficient and highly reliable while keeping delays low. In this paper we present hardware and software architecture for BAN and also we offer reliable communication and data aggregation.

Keywords: *Wireless Body Area Network, BAN, Neural Network, Interrupt*.

1. Introduction

Nowadays, one of the major applications of wireless sensor networks is environmental monitoring. In these networks, an abundance of sensors is scattered around to collect and retrieve environmental data. A new use of sensor networks can be found in the area of wearable health monitoring. Carefully placing sensors on the human body and wirelessly connecting them to monitor physiological parameters like heartbeat, body temperature, motion et cetera is a promising evolution. This system can reduce the enormous costs of patients in hospitals as monitoring can occur real-time, over a longer period and at home [1, 2]. This type of network is called a Wireless Body Area Network (WBAN) or Wireless Body Sensor Network (WBSN) [3,5,7]. A WBAN consists of several sensors and possibly actuators equipped with a radio interface. Each WBAN has a sink or personal server such as a PDA[4], that receives all information from the sensors and provides an interface towards other networks or medical staff. Connecting health monitoring sensors wirelessly improves comfort for patients but induces a number of technical challenges like coping with mobility and the need for increased reliability. An important requirement in WBANs is the energy efficiency of the system. The sensors placed on the body only have limited battery capacity or can scavenge only a limited amount of energy from their environment [6,8]. In this paper we offer Reliable communication in wireless body area sensor network for health monitoring. We organized rest of the paper section2 about architecture and section3 we offered a way for error detection, section4 explanation about communication and the end section5 is conclusion.

2. Architecture

Was designed so that the sensor nodes that are small and use the batteries so that their lives for a long time. Nodes crude weak signals from the human body are collected. The most common physiological signals, (Sp) pulse rate, respiration rate, spirometry, ECG, body temperature, blood gas levels, cardiac output, blood pressure.

Methods BAN this case acts in the human body sensors (with HUB local) player is the least disturbance for the individual to assume sensors along the blood flow in the human body vital signs his measurements are always these symptoms their control and in case of any problem in the system of human body in different ways that information to emergency centers after appropriate physician information necessary measures to person and gives the desired information as possible from the problems caused by this disorder avoids.

2.1 Architecture (for cluster head sensor)

Ordering the front view of the sensor inside the body if the body is one area such as a cluster area and imagine for a CH Cluster let out a task with environmental data collection and programs to run interrupt. Figure1 explains ordering of sensors in body.

2.2 **Architecture offer**

The architecture under a hardware interrupt occurs (figure 2). When an interrupt occurs (low level) under the program runs

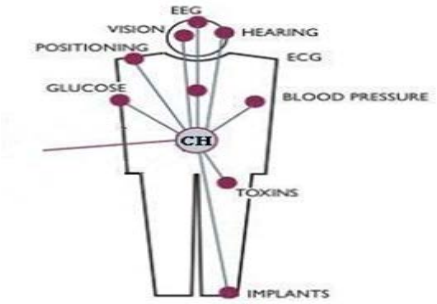
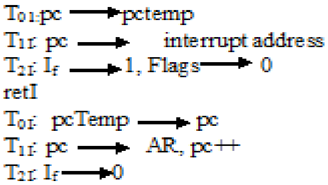


Fig. 1. Body Area Network



T_{01}:pc → pctemp
T_{1I}: pc → interrupt address
T_{2I}: I_f → 1, Flags → 0
retI
T_{0I}: pcTemp → pc
T_{1I}: pc → AR, pc++
T_{2I}: I_f → 0

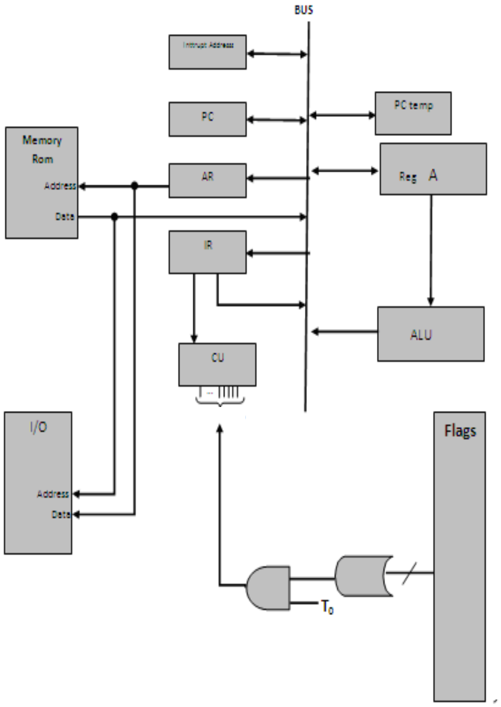


Fig. 2. Architecture

When an interrupt occurs under the program interrupt (high level) is performed (described later in different scenarios will be.) First, CH, data is sent with the normal template and immediately (after the default 100_{ms}) will go to your inbox if you did not receive a message to the following steps. First, the second data format to be sent to other cell phone I do not give two seconds if you did not answer (ACK) data format and frequency of the third device(phone) sends to the default 20 seconds to wait and interrupt is removed (RetI) Now if you interrupt the program run out again in a state of emergency occurs interrupt instruction interrupt occurs again running this operation to be performed when a state of emergency is removed or power to all sensors (sensor up their efforts person to survive will do).

$time_0 = 24_h$
$time_{01} = 1_h$
$time_1 = 100_{ms}$
$time_2 = 2_s$
$time_3 = 20_s$

A state of emergency occurs interrupt occurs

Interrupt:

Send (format data1)
Standby ($time_1$)
If (check inbox == false)

```
(
        Send (format data2)
        Standby ($time_2$)
)
Else
(If (check inbox == true)
            (Empty (inbox)
                        Standby ($time_3$)
                        Return interrupt
            )
        Else
            (Send (format data3)
Standby ($time_3$)
Return interrupt
        )
  )
```

3. Error detection

BAN ideas in this article that acts this way in all parts of the human body has been used sensors.
Here is a problem that may occur is a condition that the sensor is defective. Data from the body feels is not valid.

3.1. Error detection:

in each sensor, we built a neural network. Each sensor data in addition to those who feel their body has data of its neighboring sensors will also receive Function neural network training with the sensor data received and produced a data is that the sense data with data by the sensor is relevant to compare these two data error rate is calculated with the area defined as the maximum possible error is calculated, comparison is done if the error rate obtained is less than the amount defined by the sense data is correctly declared and will be sent to CH. For the sensor itself with the changing environment to adapt its education and only stage production is not the sensor, the neural network re-sense data correctly with the existing training and this cause we update training neural networks and sensors will lead to increase the reliability of neural network will be built. But otherwise, if the error rate range is defined more by a sense of false data is detected The first approach that, the actual range for each sensor in the human body that may happen and if we define sensor is faulty sense of data that is outside this range is timer: The timer mode, the sensor will begin to countdown if this period (until the overflow timer has not happened) data generated with data by neural network prediction is contrary it can be concluded that the sensor is disabled (After this period, individual data base based on the desired message that one of your sensor sends damaged). If before the timer overflow occurs sense data with predicted data is equal to the primary mode timer returns and is diagnosed, the sensor is not defective and the immediate effect of these uncertainties there is a contradiction.
Each sensor as Figure (3) shows includes the following sections.

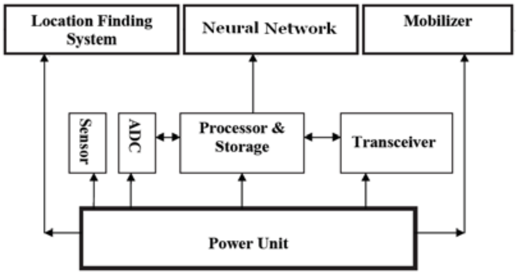


Fig. 3 Interior Sensor.

Performance of neural networks is that sensors within his duty is to determine the status of sensors (for example, send a zero element means conflict between sense data and data predicted by neural network is a means to send the same data is two. Posts by 1111101111 CH in the Status field sensors means that there are 10 sensors that sense data by the fifth sensor data generated neural network with the same sensor is different. Position Sensor in place bit location in the body determines the sensor can be that their data is important to have a bit space is valuable in this case, if the sensor is more important from the database can be displayed with high priority. the range data that each sensor can send the number of bits that send data for each sensor will be compared will be different), and vital signs that are sent with authentication.

When the cluster head receives data using a threshold to detect conditions normal or abnormal. In the normal case production of all sensor data in the threshold is an unusual case but the two conditions occur eg If less than three sensors in the threshold of any CH situation semi-critical diagnosis and a message to all broadcast to the other sensors to based on that they data to CH to send and CH after receiving their data to send the database. In critical case condition for more than three sensors placed in the threshold in this case for CH when all the broadcast message does not die in a state of crisis because the sensor data directly to other emergency and need not be informed.

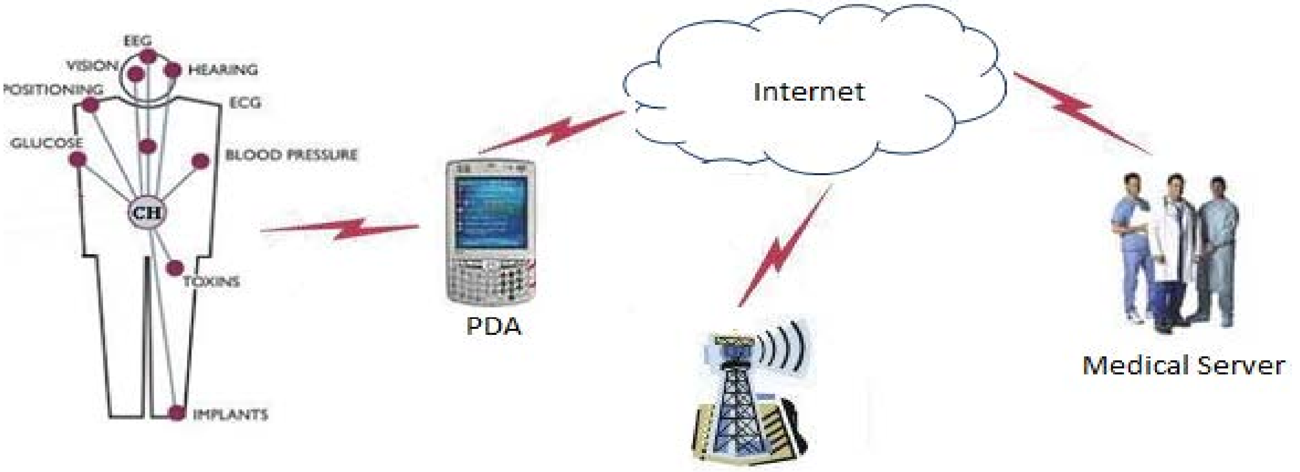


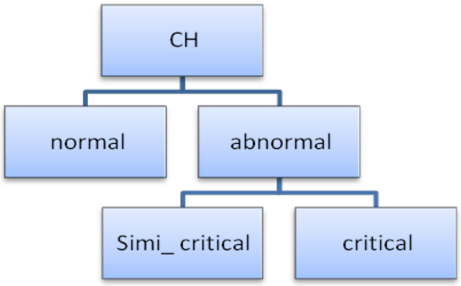
Fig. 4. General communication in BAN

4. We offer about BAN's communication

We show in figure 4 general communication in BAN and rest of the paper we explain about detail communication.

4.1. First case

One of the duties of CH, is that the individual sensors with mobile phones are considered in connection. Their relation to the way that CH vital signs such as temperature, pulse, blood pressure, blood sugar ... Humans to continuously control the body's vital signs, and if normal body's vital signs have been interrupted and is not happened in critical condition a long interval (eg once a day) by short-wavelength waves to phone and phone the person sends this information to the person via the Internet or through any other database sends personal information is stored in one file. If medical advice regarding this situation it was in the database as a phone message and sends the person himself can decide to act or not recommend it (freedom of the person he has not been forced to do work there no act.) insist on watching the arrival of these cases no information at the time specified does not exist if it did not have the phone does not delete the queue and send the information in specified intervals ($time_{01} < time_{0}$) sends. Figure 5 shows this case.



When entering the human body sensor in the database CH into a series of information that are unique for each person (for example fingerprints.) Main program when the body normally has repeatedly runs. (State of emergency has not occurred) and the program routinely done.

Data structure in the first case is as follows:

fingerprints	CRC

First abnormal format

Fingerprint	sensor status	amounts of undesirable events	Tail

First normal format

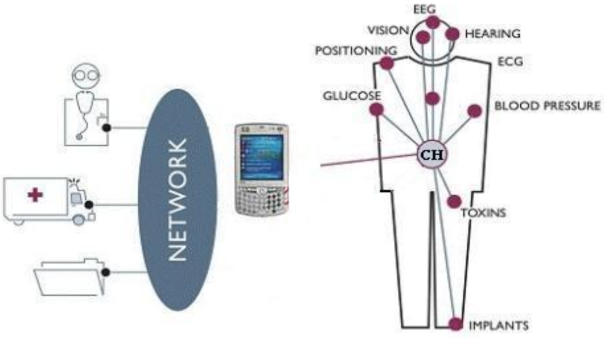


Fig. 5 . First case

Program runs in a normal situation:

```
While (the state of emergency has never occurred)
(Process (data from sensors)
     if (time system ==$time_0$)
    (
        Send (format data1)
        Send queue (format data1)
    )
    If (don't send & time system ==$time_{01}$)
    (
       If (inbox ~ = ACK)
        (
           Send queue
        )
      Else
       (
           Remove (queue)   )))
```

We can run program faster in normal times $time_0$ and $time_{01}$ If the interrupt hardware to draw until we spend more time processing CH is processing information received from sensors.

3.2. Second case

Now we consider the situation when a person is difficult, there is acute situation (in this case also interrupting the program runs continuously in the second stage of the program ends) and cell phone person is not available for this Send CH acute situation to the health care centers or emergency to Using GPRS phone and the person's status by individual service person to do.

In this case, because we are in the residential area, if the CH , cell phone does not find the person. In this case, data to be released to all broadcasting for mobile phones that are around this person (eg a radius of 40 m) received (in this section assumes that they receive) and data to send that information by health care centers or emergency situations can be patient position using other people around the individual patient (eg, radius 40 m) to find Patient records by physicians, the individual decisions are correct. The figure 6 explains this case well. (Using the Document in the database).The latter data structure is as follows: Because most likely person in the event of such a case, and because residential energy consumption in this mode is lower for the sensor to use less energy to data instruction the above Such as Bluetooth to nearby devices will send. (Energy consumption for the sensor to the mobile phone is more important).

fingerprint	CRC

Second abnormal format

bits emergency	Fingerprint	amounts of undesirable events	Tail

Second normal format

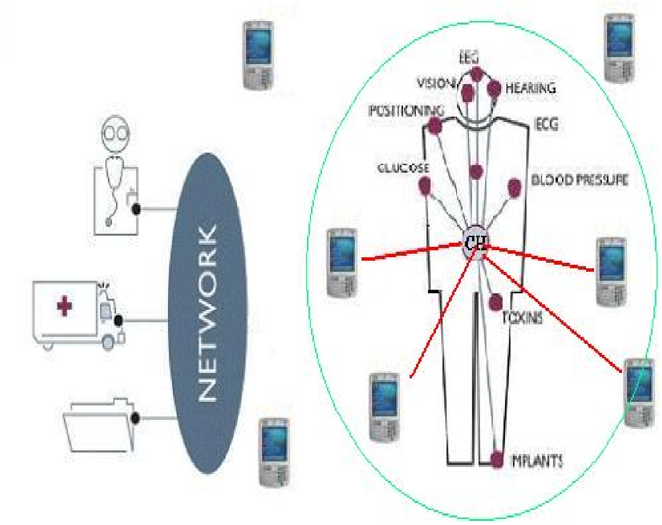


Fig .6. Second case

4.2.1. Bit of emergency

In this case, sensors that measure to submit such data to be nearby when the cell phones receive it without processing;

emergency bits removed and data downloaded to send to emergency centers and simultaneously send CH ACK. If CH does not receive answer the third case occurs. In this case, program execution stops at the second stage and then steps other cell phone will be sent. In fact, we have same interruption all cases.

4.3. Third case

Now consider the situation that we are not residential. (Eg, forest or mountains and desert interruption occurs ...) The individual is difficult. Acute situation occur and vital signs in the person gradually goes away and the person is not available for cell phones to CH this critical situation. This crisis will send to the emergency. We show this case in figure 7.In this case, because CH cell phone the person will not be the first CH data to be released to broadcast to all phones that are around this person (eg a radius of 40 m) and found to give and emergency center to inform (second case) CH because no answer from any other cell phone based on information received does not send its cluster head and decided to send the data with the same wavelength and energy production to phone and data Send to a database and database using position communications satellites or cellular towers to the nearest service center (emergency) to refer to individual service done[12,13]. (Because every person has a fingerprint data arrival is clear that the data belongs to whom)

Data structure in the third case as follows:

Fingerprint	CRC

Third abnormal format

fingerprint	amounts of undesirable events	Tail

Third normal format

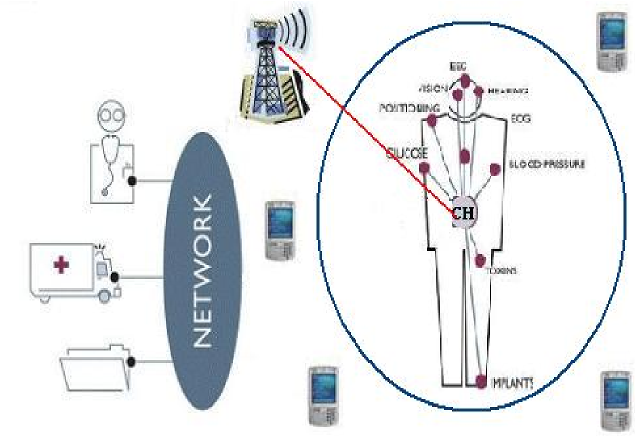


Fig. 7. Third case

5. Conclusions

Because each sensor depends on the performance of the neural network and neural network performance in such a way that its performance over time, which sees increased education and training neural networks in this paper continues and over time this leads to increased reliability neural network with time is the result of increased reliability is BAN.

In this article, as far as we are energy sensors using wavelength range, we did a short addition to informing the border have given one hundred percent Vine increases reliability is as stable relationship with Given that communication via phone and through its pessimistic mode, the sensor is done. In addition, all decisions about that person's recommendations.

References

[1] A. Hadar and S. Mahadevan, “Probability, Reliability and Statistical methods in Engineering Design”, John Wiley and Sons Inc, 2000.

[2] CodeBlue: Wireless Sensors for Medical Care, Available: http://fiji.eecs.harvard.edu/CodeBlue

[3] Q. Wang, M. Hempstead, W. Yang, “A Realistic Power Consumption Model for Wireless Sensor Network Devices,” 3rd Annual IEEE Communications Society on Sensor and Ad hoc Communications and Networks, pages 286-295, Sept. 2006

[4] S. Kellner, M. Pink, D. Meier and E.O Blass: "Towards a Realistic Energy Model for Wireless Sensor Networks," Fifth Annual Conference on Wireless on Demand Network Systems and Services, pages 97-100, Jan. 2008

[5] C. Bettstetter. "Smooth is Better than Sharp: A Random Mobility Model for Simulation of Wireless Networks," Proceedings of the 4th ACM International Workshop on Modeling, Analysis, and Simulation of Wireless and Mobile Systems, pages 19 -27, July 2001.

[6] E. Belding-Royer and C. Perkins: "Evolution and future directions of the ad hoc on-demand distance-vector routing protocol". Ad hoc Networks Journal, Vol. 1 no. 1, pages. 125-150, July 2003.

[7] A. Sobeih and J. C. Hou. : "A Simulation Framework for Sensor Networks in JSim". Technical Report UIUCDCS-R-2003-2386, November 2003.

[8] Chipcon CC2420 Zigbee/IEEE 802.15.4 RF Transceiver, Available: www.ti.com

Saeid Bahanfar received the B.Sc. degree in Computer Software Engineering from Payam Noor University (PNU), Tabriz branch, Iran in 2008._ Currently, he is a M.Sc. student of Computer System Architecture in Islamic Azad University, Tabriz branch, Iran. His research interests include Residue Number System and VLSI Design, wireless sensor network, Neural network.

Ladan Darouagarn was born in Tabriz, Iran, on May 29, 1983. She received the B.Sc. degrees from University of Shabestar (Shabestar, Iran) and M.S.E. student in Islamic Azad University, Tabriz Branch in 2011. Her research interests are in the data aggregation in wireless sensor network. She is a member of Young Researchers Club.

Helya Kousha received her B.Sc. in Computer Software Engineering from Islamic Azad University, Shabestar branch, Iran in 2008. Currently, she is a M.Sc. student of Computer System Architecture in Islamic Azad University, Tabriz branch, Iran. Her main research interests include Computer Arithmetic, Residue Number System, wireless sensor network.

Shahram Babaei He is currently DR. student in Department of Computer Engineering of Science and Research Branch of Islamic Azad University, Tehran, Iran. His research interests are Computer Arithmetic ,Reliability in wireless sensor network, Cryptography, Network Security.

A Review of Clustering Techniques Based on Machine learning Approach in Intrusion Detection Systems

Ala' Yaseen Ibrahim Shakhatreh [1], Kamalrulnizam Abu Bakar [2]

[1,2] **Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia**
81310, Johor Bahru, Malaysia

Abstract

False alarm rate and detection accuracy are still challenging issues that are not completely solved yet in the field of Anomaly based Intrusion Detection System (AIDS). The reasons behind these issues vary according to the algorithm and the dataset used to train the IDS. Consequently, dealing with high dimensional data requires an efficient data reduction technique that considerably reduces the dimensionality without any substantial loss in the important features. However, the excessive reduction of features will lead to model some intrusive patterns similarly as normal ones. Indeed, this will result in misclassifications that will increase false negative rate, which degrades the accuracy of detection. This paper concludes many clustering techniques that were previously proposed to solve the inherent IDS problems. Where, the clustering techniques involved in three general aspects namely: data preprocessing, anomaly detection, and data projection/alarm filtering. Eventually, recommendations for future researches followed by the conclusion are depicted at the end of this paper.

Keywords: *Intrusion Detection System, Clustering Techniques, Unsupervised Learning, Detection Rate, False Alarm Rate, Dataset, LVQ, SOM.*

1. Introduction

The nowadays internet has expanded without limits, as many systems moved online to gain the profit of internet marketing. Consequently, the number of security incidents has exploded. Thus, internet security became a growing concern that led many security research oriented organizations to conduct studies to provide an acceptable level of protection against intruders [1]. Intrusion Detection Systems became a necessary complement to the traditional firewalls to ensure data integrity, confidentiality and availability for data transmitted over the network. Relatively, Intrusion Detection Systems are categorized based on two approaches: misuse and anomaly based detection. Misuse is an efficient way to detect known attacks that have known hard coded signatures stored in the signature list. However, any simple variation from the listed signatures will lead to consider such an attack as a legitimate request leading to a high false negative rate. One fact about misuse approach is its low false positive rate, due to its pattern matching techniques with the signature list. However, misuse approach has failed in detecting unknown and 0-day attacks. Thus, the knowledge base must be updated frequently and manually.

On the other hand, anomaly based approach depends on establishing normal profile usage and any violation of that normal profile will be considered as an anomalous request. Accordingly, anomaly approach outperforms misuse approach in terms of detection capability of novel and unknown attacks without any advance knowledge which indicates lower false negative rate. Although anomaly detection approach uses machine learning techniques (supervised & unsupervised) to successfully identify unseen attacks, but these techniques tend to generate high false positive rate due to the high dimensionality of datasets used in the training process. Despite anomaly detection approach has low false negative rate, but it is still prone to have false negative alarms because some attacks can be conducted in more than one way, which raises the issue of modeling and the generality of attacks that are not completely solved yet.

In this study our categorization criteria for the connectionist models used in IDSs is based on the functionality of algorithms used. The first group discusses the algorithms used for data preprocessing, the second one discusses the algorithms used for detection process, and the last one is for algorithms used for data projection and alarm filtering. The rest of this paper is organized as follows: section 2.0 shows critical IDSs issues, section 3.0 discusses the variety of clustering algorithms used in IDSs, section 4.0 explores the future trend of IDS, and section 5.0 depicts the conclusion of this paper.

2. Intrusion Detection Systems Issues

All previously proposed solutions mentioned in section 3.0 have addressed the issues or part of the issues that are addressed in this section, attempting to overcome specific limitations in the field of Intrusion Detection System. In general, anomaly based techniques rely on two assumptions: First, the number of normal patterns outweighs the number of intrusive ones; second, the intrusive patterns are qualitatively different from normal patterns [2]. As a result, when using any clustering algorithm to cluster a dataset, some normal patterns mistakenly fall into an anomalous cluster which generates false positive alarms [3].

The problem of increased false negative alarms is that some attack patterns are incorrectly mapped to a normal cluster which contains normal patterns. On the other hand, the problem of false positive alarms is originated from the fact of some normal patterns are incorrectly mapped to a cluster containing anomalous patterns and labeled as anomalous cluster. Furthermore, a key factor in this problem depends on how unsupervised learning techniques adapt with data inputs through its topology, and how they conduct the training process to shape the final topology. Moreover, preprocessing data inputs before training has strong impact on the classification results, where omitting any important feature may affect the detection capabilities and leads to high false alarm rate [3].

On the other hand, an intrusive pattern may fall into a legitimate cluster which leads to a false negative alarm. The reasons behind this misclassification can be the nature of architecture/topology of the clustering algorithm itself and the values of its parameters provided. Additionally, taking into account the first assumption, if the dataset contains considerable amounts of attacking patterns, the clustering technique will consider the cluster in which anomalous patterns fall in as a legitimate cluster. In other words, the IDS can be trained to accept attacks as legitimates [4]. Moreover, polymorphic attacks are still a challenging obstacle to be solved by anomaly approach, because, many attacks are modeled in similar way to normal patterns, indeed this requires dealing with the generality of attacks.

Consequently, this concern will fire up the trade-off between the detection rate and false alarm rate [5]. Furthermore, attacks could not be distinguished from normal patterns because most of critical features in the packet headers which may lead to identify attacks are not utilized.

3.0 Cluster based IDSs

Clustering techniques are the most appropriate choice for dimensionality reduction purposes, especially when a huge multivariate dataset is involved in the training process. For instance, let’s take the well-known Self-Organizing Map (SOM) proposed by [6] as a data dimensionality reduction technique that was used either for anomaly detection or data preprocessing in the subsequent literatures in this paper. The high capability of SOM in transforming high dimensional input space onto a very low dimensional neuron space (one or two topological maps) made it preferred for dimensionality reduction purposes. Yet another important feature in SOM and some other clustering techniques is that the ability of identifying outliers presented in the dataset during training phase [7]. Relatively, [8] described SOM as the best choice for dimensionality reduction due to its hard-competitive training approach in the clustering process, which robustly allows dividing the data input to certain number of classes. More detailed information about SOM can be found in [9], [7], and [10].

In general, there are three main aspects in which clustering algorithms can cope with as follows: First, data preprocessing. Second, anomaly detection. Third, data projection and alarms filtering. As our categorization criteria will be based on these three aspects.

3.1 Clustering Algorithms for Data Preprocessing

In [11], Labib and Vemuri assumed that that real-time processing and performance can be achieved using clustering technique specifically by SOM algorithm, by implementing simpler design.

3.1.1 Real-time Data Preprocessing Using Self-Organizing Map

According to the above mentioned assumption in [11], SOM was used in [12] to support online preprocessing for data preprocessing stage. The Distribution Gravity Center (DGC) was used for normalization in the preprocessing stage due to its improvements of the firing behavior of SOM instead of the Euclidean normalization method. Furthermore, the source and destination addresses in the packet headers captured from different protocols were translated into octet vectors to reduce the redundancy in the preprocessed patterns, ignoring the upper two redundant octets in all vectors. Thus, the lowest two octets along with the encoded protocol are involved in the preprocessing process. Moreover, for better pattern clustering, Kohonen Random initialization function was selected over others. The results show better detection capability and the ability to preprocess data in real-time comparing with the original SOM.

3.1.2 IDS based on Self-Organizing Map and K-mean algorithms

S-K algorithm was proposed in [7], to provide better detection and lower false positive arte by combining SOM and K-mean algorithms. Basically, SOM preprocessed the data to produce a number clusters with centers. On the other hand, K-mean algorithm is applied to refine the final results of SOM topology by eliminating grays clusters and remaining black and white ones. The results show that the new algorithm has achieved 92% detection rate and 35% false rate. However, these results were obtained based on testing the algorithm using up to three types of attacks only.

3.1.3 Feature Reduction Using Principal Component Analysis (PCA) Algorithm

The current IDS proposals emphasize on reducing the number of features as inputs to the neural model in order to maintain the performance and the efficiency. However, omitting some feature could result in modeling such an attack similarly to normal pattern, which consequently increases false alarm rate and reduces detection rate. Principal Component Analysis (PCA) algorithm was used as a compacting technique in [13], after all significant features were analyzed without any substantial loss in the information. Therefore, PCA is responsible for keeping the necessary amount of sufficient data for classification and maintaining the performance of the classifier by keeping the number necessary data at its minimum without affecting the effectiveness. As a result, the input vector for every TCP/IP packet to the neural model contains 20 inputs (extracted from 419 inputs of the original TCP/IP inputs) which made it 438 times faster than the original one.

As mentioned before, involving all features can guarantee better and accurate detection, but with more shortcomings on the real-time efficiency. Therefore, in [14] the Principal Component Analysis algorithm (PCA) was used as a feature selection algorithm in order to increase the detection rate along with the accuracy and to decrease the total complexity by reducing the dimensionality of the sample inputs. Moreover SOM was used for clustering and anomaly detection. For the sake of better evaluation, five new attack types were added to the typical attacks presented in the obtained 10% of KDD Cup 1999 as a sample dataset. The results of the proposed MPCA-MSOM algorithm shows better detection rate and lower false positive rate, 97.0% and 2.2% respectively comparing with the original SOM, K-NN, and RoughSet algorithms. As a result, the MPCA-MSOM algorithm not only proved to have better detection rate and false positive rate, but provided better attack classification.

3.1.4 Prior Knowledge for Better Traffic Characterization Using SOM and Learning Vector Quantization (LVQ)

In [5], eleven SOMs, one for every attribute selected to validate the variation of each attribute of the packet headers separately over time window, aiming to identify the main features of every attribute. While the second layer consisting of 6 by 6 SOM and LVQ assigned to each SOM in the first layer to correlate the first layer information between the attributes and classifies them among three classes: Normal, Attack or Indefinite. The last layer decides whether the vectors in the indefinite class are Normal or Attack using SOM. Furthermore, the LVQ network which is assigned to each SOM in the second level is to make the final classification of Indefinite to either Normal or Attack. The results show that the detection rate decreased by 19% which rates below comparing with: clustering, K-NN, SVM, and SOM Hierarchy, because user-to-root attack is difficult to model using the characteristics of TCP/IP traffic only.

3.1.5 Detecting 0-Day Attacks Using CMLHL Connectionist Model

Utilizing CMLHL connectionist model among multi-agent system in [15] to enhance pattern classification, for the sake of detecting 0-day attacks and other attacks conducted through SNMP protocol. Additionally, the architecture consists of one central IDS agent for anomaly detection using the connectionist model served by several sniffer agents distributed among all network segments sniffing and preprocessing packet headers captured. On important aspect of this approach is that it shows the temporal relationship between packets within time dimension. However, this architecture creates a bottleneck situation due to the huge payloads flocking from sniffer agents to the IDS agent, which requires the IDS agent to be on a huge calculus power machine. Moreover, the scope of attacks is related to SNMP attacks only.

3.1.6 Modeling Real-World Traffic to Generate Syntactic Dataset

A framework was developed in [16], to generate synthetic traffic based on HTTP protocol to measure the improvements of the generated synthetic traffic over the 10% of KDD Cup 1999 (containing HTTP connections only) and compare the results with the results of real-world traffic when testing them using K-mean and SOM based IDSs. Moreover, the traffic was generated in tcpdump format, and BRO network analyzer was customized to extract 41 features. However, only 6 features were involved in this testing namely: duration of the connection; protocol; service; connection status; total bytes sent to destination host; total bytes sent to source host. After testing the three datasets using the K-mean clustering algorithm, it can be concluded that the generated synthetic dataset has more similarities to the real-world traffic (generated from Faculty of Computer Science server Locutus in Dalhousie University) than KDD Cup 1999.

On the other hand, when testing the datasets using SOM based IDS developed by [15], by applying two hierarchy levels of SOM. At the first level each feature of the six features is assigned to a SOM to be trained aiming to encode temporal relationships among the features. The second level, the information from the first level is combined to be represented and labeled. The results concluded that the SOM trained on the synthetic dataset shows improvements and more similarities to the SOM trained on the real-world dataset than the one that trained on KDD99.

3.2 Clustering Algorithms for Anomaly Detection

Supervised and unsupervised clustering algorithms were utilized and improved to enhance detection capabilities and false alarm rate.

3.2.1 Combining SOM Algorithms for Scalable IDS

In order to avoid the high complexity of the training process of the original SOM, the most proper SOM type is selected and assigned to network node in [18], that in order to cope with real-time requirements by reducing the training cost. The selection of SOM type is based on the following criteria: scalable SOM is used if the number of zero-elements is greater than 40%, and there is no memory limitation. Scalable SOM with compressed vectors is used if the memory limitation matters. If the number of features is lower than 10, the original SOM is used. Otherwise, if it is very high more than 500 or the memory resources are very limited or the visualization is not needed, then HSOM is used because it is fast. Additionally, if there is no memory limitation and the CPU has limited resources, then GHSOM is used. Moreover, if the importance of node is very high, then the original SOM is involved. Furthermore, fast winner search technique is used whenever real-time processing is important. As a result, the optimal training cost can be achieved based on these criteria.

3.2.2 Multi-agent IDS based on Clustering Algorithm

Yet another approach to enhance detection capabilities through multi-agent system or distributed IDS as applied in [15], [19], and [20]. Thus, enhancing pattern classification in [15] through CMLHL connectionist model is described in section 3.1.5. While in [19], the IDS is composed of distributed agents (one for each network node) administrated by administrator agent which uses clustering technique based on new growing Self-Organizing Map model to provide flexibility and modularity in detecting anomalies. Moreover, the distributed agents detect the anomalies based on the shared knowledge in the administrator agent. After training and testing the system using wide scope of attacks, about 22 attack types in addition to normal patterns presented in KDD Cup 1999, the results showed 90.79% detection rate which is lower than other works due to wider scope of attacks and features involved.

Combining Case-Based Reasoning (CBR) and neural network through multi-agent system in [20], to increase the detection and projection capabilities of SNMP related anomalies. The architecture consists of six evolving agents that can dynamically learn and adapt using the neural model. The analyzer agent is a CBR-BDI agent that applies neural model within its adaptation stage for analyzing the preprocessed segments to allow the projection of network traffic. Furthermore, several neural models namely: PCA, CCA, MLHL, and CMLHL were applied on a dataset obtained from SNMP traffic using 5 features for comparison reasons. As a result, CMLHL model found to have the best projection and attack identification comparing with MLHL, PCA, and CCA models. Moreover, with respect to CCA, the best results were depicted based on Standardized Euclidean Distance.

3.2.3 IDS based on Dynamic Self-Organizing Map DSOM

Anomaly clusters can be identified by the cluster of normal ratio. Thus, in [21], [22], and [2], improving pattern classification is through the use of dynamic SOM (DSOM) with other clustering algorithms to make the detection independent of the centers of clusters. Apart from traditional clustering by which the accuracy of detection is degraded by the use of simple distance metric, the growing DSOM and Ant Colony Optimization algorithm (ACO) can control clustering efficiency in [21]. Additionally, in the detection stage, Posteriori probabilities make it much independent to increase the efficiency in detecting unknown attacks. Then, after initializing ants, each ant selects an object randomly, and picks it up or moves or drops it based on the probability of each action in DSOM. In the second stage, clusters are gathered from DSOM's output preparing for labeling the obtained clusters. As the detection stage is based on Bayes theorem due to its low fault rate. Eventually, using KDD dataset for testing, the experimental results show better detection rate and false positive rate comparing with K-NN and SVM based IDS.

On the other hand, the same DSOM was used in [22] followed by Swarm Intelligence (SI) clustering for the same purpose in [21], but in hope to more accurate detection classifier to enhances the detection rate and false positive rate. The results of this approach show better performance in detecting intrusions comparing with LGP, SVM, KNN, and DT. As a result, from these two works, we can conclude that Swarm Intelligence cope better with Dynamic SOM than ACO in quality of clustering and intrusion detection accuracy.

Additionally, after obtaining clusters by Dynamic SOM (DSOM) in [2], Bayesian classification algorithm is optimized as the detection algorithm, which has the least fault rate among other classification algorithms. Furthermore, Bayes theorem makes the detection process independent of the center of clusters, which increase the accuracy of detection. Using dataset D obtained from KDD Cup 1999 which consists of 1% to 1.5% intrusions, and 98.5% to 99% normal patterns for testing and evaluating the proposed IDS. Consequently, the results show higher detection rate and lower false positive rate comparing with Cluster, K-NN, and SVM based IDSs.

3.2.4 Hierarchical Clustering Based IDS

Hierarchical cluster techniques were introduced in [10], [23], [24], [25], and [26] to address the main limitations of traditional clustering techniques. Where, the static architecture is imposed by establishing a fixed topology in advance before the training stage. Moreover, the topology maintains many empty and unnecessary clusters which degrade the efficiency by increasing the complexity of the IDS. Thus, the input vectors are not faithfully represented in the traditional clustering topology, which results in lower detection accuracy and more false alarm rate.

However, in [10], Growing Hierarchical SOM (GHSOM) is optimized for better classification. Considerably, GHSOM consists of several SOMs arranged in layers growing during the training process along with the number of layers and neurons of maps to automatically adapt with data inputs can faithfully represent input vectors. Moreover, a metric based on entropy for symbolic values together with numerical values (by Euclidean distance) is presented in GHSOM in order to involve three pivotal symbolic features namely: protocol type, service, and flag of status in the training stage. After training the model using KDD Cup 1999 dataset, the results of this GHSOM were compared with SOM and K-Map algorithms. The experiments showed detection rate: 99.98, 81.85, and 99.63% respectively and in terms of false positive rate: 3.03, 0.03, and 0.34% respectively, which indicates better detection capabilities with an acceptable FPR.

The growing hierarchical self-organization graph (GHSOG) is proposed in [23], to overcome the limitations of static topology and the lack of representation of hierarchical relations between data inputs in SOM. On the other hand, GHSOM has also some limitations related to the static topology, where each map is initiated with 2x2 neurons, forcing 2-D rectangular grid of map adding many neurons without necessity, taking the map far from the optimal number of neurons. Consequently, GHSOG is based on establishing map topology according to data inputs faithfully, by reflecting data inputs as faithfully as possible. KDD Cup99 was used for training and evaluating the model. Moreover, to avoid preprocessing traffic data in consecutive quantitative way, each qualitative feature is replaced with binary vector consisting of many binary features, allowing using the Euclidean distance function. By this, the number of features increased from 41 to 118. The results show lower detection rate about 90.68% and more FPR comparing with K-Map, SOM, and SOM (DoS). However, the mentioned works used only 3 attack types from KDD Cup99, while in this work 38 attack types were used, where 15 are new and unknown attacks. Furthermore, this model can cope with real-time IDS due to its lower complexity by utilizing lower number of neurons than other works.

Enhanced SVM is used to provide better performance and generalization accuracy in [24]. In this work hierarchical clustering analysis is used through Dynamic Growing Self-Organizing Tree (DGSOT) to cluster huge dataset efficiently and to overcome the limitation of time consuming training phase of SVM. Considerably, DGSOT helps in SVM training by finding the best qualified boundary points between two classes. This work has contributed in: First, reducing SVM training time. Second, the enhanced SVM is proved to be faster than the original. Third, in terms of false positive FP and false negative FN rates this approach outperforms random selection and Rocchio Bundling on a benchmark dataset. After using DARPA dataset, involving DOS, U2R, R2L, and Probe attacking patterns, the results show enhancements over Random Selection, pure SVM, and SVM + Rocchio Bundling in terms of detection rate and FPR. However, this approach still shows low accuracy for detecting U2R and R2L attacks 23% and 43% respectively, which indicates high false alarm rates for these attacks.

The evolution of clustering techniques is still at its peak, leading to innovate new elegant techniques that take the advantages of more than one clustering algorithm. This approach can be seen in this paper through the evolution of Self-Organizing Map algorithm (SOM). The new upgraded SOM models led to better detection rate and lower false alarms. Moreover, real-time efficiency was taken into account in growing hierarchical related models as applied in [10], [23], and [24]. On the other hand, Growing Hierarchical Recurrent SOM (GH-RSOM) was used for efficient clustering in phoneme recognition in [25]. The contribution in this work is to make a hierarchical model that composed of independent RSOMs; each one is allowed to grow during the unsupervised learning process until the quality of data representation is met. Moreover,

the best matching unit is selected through the difference vector defined by each map. And the adaptation of the weight of the map is defined by the difference vector as well. Eventually, comparing the proposed model with GHSOM model, the proposed algorithm (GH_RSOM) provides better classification rates of phoneme recognition. From our perspective, this model is applicable for intrusion detection system to provide better attack classification.
A new Probabilistic Self-Organizing Graph (PSOG) algorithm is proposed in [26], to provide better classification capabilities. And the topology of Self-organizing neural model is adapted to reflect the internal structure of inputs distribution rather than being fixed during the learning phase as applied in other works. Moreover, each unit of the resulted self-organizing graph is a mixture component of Gaussians (MoG). Furthermore, the corresponding update equations are obtained from stochastic approximation framework, as it is used to learn both mixture and the topology. As each probabilistic mixture of multivariate Gaussian components is associated with a neuron or unit. For evaluation purposes, four uniform distributions were selected to show graphically how PSOG model can adapt its topology according to the structure of input distributions. From the results, PSOG shows better adaptation to its inputs and can perform classification better than other static models, because other static models do not learn their topologies.

3.2.5 Novelty Detection Using Clustering Techniques

In this section, two studies were conducted to achieve novelty detection based on combining SOM with Genetic algorithm to produce GSOMS in [27], and comparing SOM-L with One-class Learning Vector Quantization (OneLVQ) in [28] in terms of classifying novel patterns during training.

A combination of SOM and Genetic algorithms is to form Genetic SOM clustering algorithm order to increase the effectiveness of training process in [27]. Moreover, the role of genetic algorithm is for training the synaptic weights of SOMs. In other words, the adjustment of the SOM synaptic weights is conducted through GA instead of traditional learning rules. Specifically, a chromosome of the GA represents possible combinations of synaptic weights of SOM neurons. In previous studies, detecting novel attacks was based on knowledge learnt from labeled data. Furthermore, detecting novel attacks requires training with new labeled data samples. However, this solution (training with new labeled data) is very expensive when dealing with huge network data. In order to evaluate the proposed mode, KDD Cup99 was used, and the 41 features of each connection were divided into 4 categories according to their data types as follows: Strings, Boolean, Count (integer), and Rate (float) types. With the optimal K representative value which equals to 40, GSOMC model achieved 80% detection rate and 1.9% false alarm rate. The low detection rate explains that some attacks presented in the KDD Cup 1999 dataset or in the real-world internet are stealthy and difficult to model using Euclidean distance measure used in many techniques.

In [28], two methods were proposed to detect novel patterns. First, presenting a scheme of SOM-L boundary to determine local threshold. Second, modifying the learning rule of one-class Learning Vector Quantization (OneLVQ) to allow one to keep codebook vectors far from novel patterns as much as possible. A key factor in this work is utilizing the novel patterns presented in the training dataset. According to [29], novel patterns in the training can be utilized to achieve high classification performance. Furthermore, novelty detection means proper generalization to characterize patterns from normal class. In the meanwhile, specialization means excluding patterns from all other classes [30]. Consequently, proper balancing between these two concepts can achieve classification performance. Unlike the original LVQ, OneLVQ is based on LVQ learning rule, but it assigns the codebooks to one class only (the normal class) rather than many classes as in the original LVQ. Accordingly, the error rate is minimized in which the codebooks are forced to be located near the normal patterns and far from novel ones. After testing the two models, OneLVQ correctly classified all three regions O1, O2, and O3 without any misclassification, outperforming the SOM-L method which failed in recognizing large part of normal patterns. As a conclusion, OneLVQ outperform SOM-L in terms of novelty detection and utilizing novel patterns during training phase.

3.3 Clustering Algorithms for Data Projection and Alarm Filtering

Traditional IDSs tend to generate huge amount of alarms during the detection stage, which exhausts system administrator by rendering a large amount of unserious alarms. Fittingly, these enormous amounts of alarms are filtered and utilized to reduce false positive rate and to increase detection accuracy in [3], [31], and [32]. A data mining technique based on Growing Hierarchical Self-Organizing Map (GHSOM) which adjusts its topology during the learning process according to the inputs data (alarms) to reduce false positive alarms and to assist system administrators in analyzing alarms generated by the IDS. The proposed algorithm aims to explore the hidden structure of alarm data, and to uncover false alarms (FP & FN) hiding in normal clusters. Considerably, a data sample consisting of 1849 data patterns including 6 web attack scenarios were tested using the proposed approach. The results show that the proposed algorithm outperforms SOM algorithm in terms of both false positives and false negatives which reduced from 15% to 4.7% and from 16% to 4% respectively.

A huge amount of alerts generated from the IDS are correlated to such categories to remove the unnecessary or unserious alerts and to make them readable for the administrators in [31]. A major limitation in the previous solutions of alert correlation is that the methods used led to increase false positives. In order to cope with this problem, in this study selected features from alerts are fed to the SOM to provide better correlation. The selected features should be able to identify the behavior of the traffic. Source/destination IP, target port, source IP of the current alert, target IP of the previous alert, sensor ID, and signature ID are capable to identify the intention of the traffic. Accordingly, if two alerts have the same six features, they will be clustered in the same neuron in SOM. Relatively, if two alerts have five similar features, they will be clustered in the vicinity of the six features neuron, and so on until two features are matched. This model can certainly help system analysts in identifying intrusions by concentrating on the groups of alerts that are relevant with each other.

A detection and prevention layer from SQL-Injection attacks through an anomaly visualization agent was proposed in [32], as a complement to an existing IDS called SCMAS. Furthermore, SCMAS has been upgraded by adding new agent (visualizer), its main functionality is to compliment the classification of this attack by improving the performance of classification. Three types of projection models were applied in the anomaly agent namely, Principal Component Analysis (PCA), Curvilinear Component Analysis, and Cooperative Maximum Likelihood Hebbian Learning (CMLHL) as their results were compared with each other. In order to test the model, SQLMAP 0.5 was used to generate malicious queries using all possible types of SQL-injection attacks. Totally, a sample of 1000 normal and malicious patterns was selected as a dataset. As a result, CCA was proved to have the best classification by grouping normal patterns separately from anomalous ones. On the contrary, PCA and CMLHL mix normal patterns with anomalous ones which increase false alarms.

4. Recommendations for Future Researches

At the early time of intrusion detection system, misuse approach was applicable due to the limited number of attacks at that time. For instance, SNORT IDS was widely and successfully applied to provide sufficient protection. However, the exponential growth of new attacks has pushed the researchers to move forward to anomaly approach, due to the insufficiency of misuse approach in detecting unknown attacks. Moreover, based on the comprehensive survey in [33], Soft Computing methods (SC) consisting of Fuzzy Logic (FL), Artificial Neural Network (ANN), Probabilistic Reasoning (PR), and Evolutionary Computing played a significant role in improving the accuracy of detection and false alarm rate through three approaches: consecutive combinations, ensemble combinations, and hybrid combinations. Furthermore, ensemble combinations of several SC methods in parallel as described in figure 1, was proved to be more robust than consecutive and hybrid combinations.

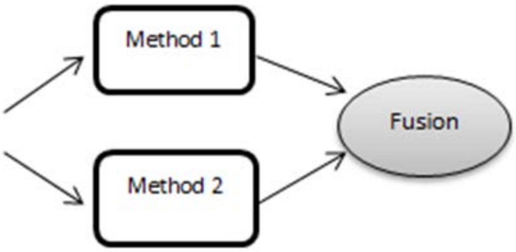


Fig. 1 an ensemble strategy.

Our recommendations for future research are represented in optimizing SC methods in ensemble approach among five stages namely: data preprocessing, features reduction, clustering, training, and classification. However, our main objective in our research is to provide applicable solution in the real-world internet by increasing the scope of attacks along with detecting novel attacks in the detection process with more accuracy. Furthermore, hierarchical clustering techniques were proved to be more robust than other techniques in terms of clustering accuracy and complexity when the number of attacks involved is higher. Relatively, fuzzy classification with rule generation based on partition of overlapping areas was proved to be the most accurate in attacks classification in [33] about 100% accuracy. Consequently, clustering and classification outputs can be more accurate if all important features of connections are involved. However, in order to avoid the high computational cost of the whole process, an efficient feature reduction technique can be optimized without any substantial loss in the important features that may lead to increase the accuracy of detection.

5. Conclusion

In this paper, we walked through the development of anomaly based intrusion detection systems during the recent years. As several supervised and unsupervised clustering techniques were optimized resulting in more elegant techniques that provided more detection accuracy and lower false alarm rate. Moreover, the newly proposed techniques tend to avoid the creation of unnecessary neurons in the training process to faithfully represent data inputs as applied in hierarchical clustering. Furthermore, this restriction in creating neurons significantly contributes in reducing the complexity of the training process and producing more accurate topologies. Since, our main concern in our research is to increase the quality of clustering and attacks classification for larger scope of attacks. Additionally, increasing the identification rate of novel patterns in the training process as well.

References

[1] A. Bashah Mat Ali, A. Yaseen Ibrahim Shakhatreh, M. Syazwan Abdullah, and J. Alostad, “SQL-injection vulnerability scanning tool for automatic creation of SQL-injection attacks,” in Procedia Computer Science, 2011, vol. 3, pp. 453-458.

[1] A. Bashah Mat Ali, A. Yaseen Ibrahim Shakhatreh, M. Syazwan Abdullah, and J. Alostad, “SQL-injection vulnerability scanning tool for automatic creation of SQL-injection attacks,” in Procedia Computer Science, 2011, vol. 3, pp. 453-458.

[2] Y. Feng, K. Wu, Z. Wu, and Z. Xiong, “Intrusion Detection Based on Dynamic Self-organizing Map Neural Network Clustering 2 Intrusion Detection Based on DSOM Clustering,” in Proceedings of ISNN 2005, pringer-Verlag Berlin Heidelberg, 2005, pp. 428-433.

[3] N. Mansour, M. I. Chehab, and A. Faour, “Filtering intrusion detection alarms,” Cluster Computing, vol. 13, no. 1, pp. 19-29, Sep. 2009.

[4] G. Giacinto, “Detection of Server-side Web Attacks,” in JMLR: Workshop and Conference Proceedings 11, 2010, vol. 11, pp. 160-166.

[5] A. Carrascal, J. Couchet, E. Ferreira, and D. Manrique, “Anomaly Detection using prior knowledge : application to TCP / IP traffic,” IFIP International Federation for Information Processing, Artificial Intelligence in Theory and Practice, vol. 217, pp. 139-148, 2006.

[6] T. Kohonen, “uni Out o t cessi s,” Biological Cybernetics, vol. 69, pp. 59-69, 1982.

[7] W. Huai-bin, Y. Hong-liang, X. Zhi-jian, and Y. Zheng, “A Clustering Algorithm Use SOM and K-Means in Intrusion Detection,” in 2010 International Conference on E-Business and E-Government, 2010, no. 2007, pp. 1281-1284.

[8] V. K. Pachghare, V. a Patole, and D. P. Kulkarni, “Self Organizing Maps to Build Intrusion Detection System,” International Journal of Computer Applications, vol. 1, no. 8, pp. 1-4, Feb. 2010.

[9] D. V. Raje, H. J. Purohit, Y. P. Badhe, S. S. Tambe, and B. D. Kulkarni, “Self-organizing maps: a tool to ascertain taxonomic relatedness based on features derived from 16S rDNA sequence.,” Journal of biosciences, vol. 35, no. 4, pp. 617-27, Dec. 2010.

[10] E. J. Palomo, E. Domínguez, R. M. Luque, and J. Muñoz, “An Intrusion Detection System Based on Hierarchical,” in Proceedings of the International Workshop on Computational Intelligence in Security for Information Systems CISIS’08, 2009, vol. 53/2009, pp. 139-146.

[11] K. Labib and R. Vemuri, “NSOM : A Real-Time Network-Based Intrusion Detection System Using Self-Organizing Maps,” Networks and Security, 2002.

[12] M. Angel and P. Del, “Towards an Intelligent Intrusion Detection System based on SOM Architectures,” Applied Sciences, pp. 1-13, 2006.

[13] I. Lorenzo-fonseca, F. Maciá-pérez, and F. J. Mora-gimeno, “Intrusion Detection Method Using Neural Networks Based on the Reduction of Characteristics,” in IWANN ’09 Proceedings of the 10th International Work-Conference on Artificial Neural Networks: Part I: Bio-Inspired Systems: Computational and Ambient Intelligence, 2009, pp. 1296-1303.

[14] J. Bai, Y. Wu, G. Wang, S. X. Yang, and W. Qiu, “A Novel Intrusion Detection Model Based on Multi-layer Self-Organizing Maps and Principal Component Analysis,” Proceedings of ISNN 2006, Springer-Verlag Berlin Heidelberg, pp. 255 - 260, 2006.

[15] E. Corchado, Á. Herrero, and J. M. Sáiz, “A FEATURE SELECTION AGENT-BASED IDS,” Symposium A Quarterly Journal In Modern Foreign Literatures, 2007.

[16] H. G. Kayac and N. Zincir-heywood, “Analysis of Three Intrusion Detection System Benchmark Datasets Using Machine,” in Proceedings of the IEEE ISI 2005, 2005, pp. 362 - 367.

[17] H. G. Kayacik, A. N. Zincir-Heywood, and M. I. Heywood, “On the capability of an SOM based intrusion detection system,” in Proceedings of the 2003 IEEE IJCNN, Portland, USA, July 2003, 2003, vol. 3, pp. 1808-1813.

[18] S. Albayrak, C. Scheel, D. Milosevic, and A. Muller, “Combining Self-Organizing Map Algorithms for Robust and Scalable Intrusion Detection,” in International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC’06), 2005, pp. 123-130.

[19] E. J. Palomo, E. Dom, R. M. Luque, and J. Mu, “A Self-Organized Multiagent System for Intrusion Detection,” in Agents and Data Mining Interaction, 2009, pp. 84-94.

[20] Á. Herrero and E. Corchado, “Agents and Neural Networks for Intrusion Detection,” International Workshop on Computational Intelligence in Security for Information Systems 2008, pp. 155-162, 2009.

[21] Y. Feng, J. Zhong, Z.-yang Xiong, C.-xiao Ye, and K.-gui Wu, “Network Anomaly Detection Based on DSOM and ACO Clustering,” in Springer-Verlag Berlin Heidelberg, Part II, LNCS 4492, 2007, pp. 947-955.

[22] Y. Feng, J. Zhong, Z.-yang Xiong, C.-xiao Ye, and K.-gui Wu, “Intrusion Detection Classifier Based on Dynamic SOM and Swarm Intelligence Clustering,” Springer Science+Business Media B.V, pp. 969-974, 2008.

[23] E. J. Palomo and D. L, “Hierarchical Graphs for Data Clustering,” in IWANN ’09 Proceedings of the 10th International Work-Conference on Artificial Neural Networks: Part I: Bio-Inspired Systems: Computational and Ambient Intelligence, 2009, pp. 432-439.

[24] L. Khan, M. Awad, and B. Thuraisingham, “A new intrusion detection system using support vector machines and hierarchical clustering,” The VLDB Journal, vol. 16, no. 4, pp. 507-521, Aug. 2006.

[25] C. Jlassi, N. Arous, and N. Ellouze, “The Growing Hierarchical Recurrent Self Organizing,” in NOLISP 2009, LNAI 5933, Springer, 2010, pp. 184-190.

[26] E. López-rubio, J. M. Ortiz-de-lazcano-lobato, and M. C. Vargas-gonzález, “Probabilistic Self-Organizing Graphs,” in IWANN 2009, Part I, LNCS 5517, 2009, pp. 180-187.

[27] C. C. Lin and M. S. Wang, “Genetic-clustering algorithm for intrusion detection system,” International Journal of Information and Computer Security, vol. 2, no. 2, p. 218, 2008.

[28] H.-joo Lee and S. Cho, “SOM-Based Novelty Detection Using Novel Data,” in Proceedings of IDEAL 2005, 2005, pp. 359-366.

[29] Y. Zhao and Z. Wang, “[Support vector data description for finding non-coding RNA gene].,” Sheng wu yi xue gong cheng xue za zhi = Journal of biomedical engineering = Shengwu yixue gongchengxue zazhi, vol. 27, no. 4, pp. 779-84, Aug. 2010.

[30] D. Fisher, “Machine Learning, Special Issue on Unsupervised Learning, 1?? (to appear),” Computer, vol. 5, pp. 1-29, 2001.

[31] M. Kumar, S. Siddique, and H. Noor, “Feature-based alert correlation in security systems using self organizing maps,” in Proceedings of SPIE, 2009, no. Id, pp. 734404-734404-7.

[32] Á. Herrero, C. I. Pinzón, E. Corchado, and J. Bajo, “Unsupervised Visualization of SQL Attacks by Means of the SCMAS Architecture,” in Trends in PAAMS, AISC 71, Springer, 2010, pp. 713-720.

[33] C. Langin and S. Rahimi, “Soft computing in intrusion detection: the state of the art,” Journal of Ambient Intelligence and Humanized Computing, vol. 1, no. 2, pp. 133-145, Apr. 2010.

Ala’ Yaseen Ibrahim Shakhatreh obtained his master degree in Information Technology from Universiti Utara Malaysia (UUM), and currently he is a PHD student in the Department of Computer Systems and Communications of Computer Science and Information Systems Faculty at the Universiti Teknologi Malaysia. His research area is in network security (Intrusion Detection System) and penetration testing. As he is supervised by Assoc. Prof. Kamalrulnizam Abu Bakar.

Kamalrulnizam Abu Bakar obtained his PhD degree from Aston University (Birmingham, UK) in 2004. Currently, he is an Associate Professor in Computer Science at Universiti Teknologi Malaysia (Malaysia) and member of the “Pervasive Computing” research group. He involves in several research projects and is the referee for many scientific journals and conferences. His specialization includes mobile and wireless computing, information security and grid computing.

Secure Geographic Routing Protocols: Issues and Approaches

[1]Mehdi sookhak,[2]Ramin Karimi,[3]NorafidaIthnin,[4]Mahboobeh Haghparast, [5]Ismail FauziISnin

[1,2,3,5]Faculty of Computer Science and Information system
University technology Malaysia
Johor, 81300, Malaysia

[4]Faculty of Library and Information Science
University of Isfahan, Iran

Abstract

In the years, routing protocols in wireless sensor networks (WSN) have been substantially investigated by researches Most state-of-the-art surveys have focused on reviewing of wireless sensor network .In this paper we review the existing secure geographic routing protocols for wireless sensor network (WSN) and also provide a qualitative comparison of them.

Keywords- *Wireless Sensor Network, Sensor, Geographic Routing.*

1. Introduction

According to great capabilities of WSNs, application of them is increasing in recent decade. But, they face to some challenges such as limitation of power, memory, CPU and etc. these issues of WSNs have a direct effects on algorithms that are designed to them because complex algorithms need much memory and CPU and they consume a great deal of energy. These extreme limitations of resource, separate WSNs from traditional networks [1]. Based on the natural features of WSNs that distinguish them from other wireless networks such as ad hoc networks, routing in WSNs has very challenges. First, establishing comprehensive structure of address for deploying of the certain number of sensor nodes is impossible. So, traditional methods based on IP address (IP-based protocols) cannot be used to wireless sensor networks. Second, almost all applications of sensor networks need to sense the flow of data from multiple sources and transfer them to a special sink that it is as opposed to communication networks. Third if multiple sensors that are deployed in the adjacency of an event create same data, the data traffic is generated that it has an important redundancy in it. Such redundancy requires to be developed by the routing protocols to make energy and bandwidth utilization better. Finally, sensor node needs an accurate resource management because the resources of sensors such as energy, power of sending packet and the storage of sensor is restricted [2].

One of the important issues in WSNs is to provide the security of sensor nodes. There are various sensor holes as sink/black holes, worm holes, Sybil attack and etc. may form in a WSN and create network topology variations which trouble the upper layer applications [3]. Among all attacks, wormhole has more significant threat; because this type of attack does not need to compromise a sensor in the network and it can create the other type of attack easily. On the other hand, using a cryptographic technic cannot prevent wormhole attack [4].

2. Security Issues and Attacks on Sensor Network Routing

Most wireless sensor networks routing protocols are not complicated and they cannot protect themselves against large range of attacks. The attacks that can effect on WSNs are belonged to one of the following categorizations: spoofed, altered, or replayed routing information, selective forwarding, sinkhole attacks, Sybil attacks, wormholes, HELLO flood attacks, acknowledgement spoofing. The descriptions of each attack are mentioned in below [5].

2.1. Spoofed, Altered, or Replayed Routing Information

The main goal of most direct attack to routing protocol is to alter or modify the information that transmitted among nodes. create routing loops, attract or repel network traffic, extend or shorten source routes, generate false error messages, partition the network, increase end-to-end latency, etc. are some side effects of spoofing, altering, or replaying routing information on sensor networks.

2.2. Selective Forwarding attack

If an attacker intercepts or refuses to transmit a certain message and either drops it or chooses an arbitrary message for sending due to stop important message, the selective forwarding attack is occurred.

This attack may be appeared in two forms. In the simplest form of this attack, adversaries try to use a malicious node for rejecting and dropping all received packets. This type of selective forwarding attack operates like a black hole.

A second type of this attack happens when an adversary modifies transmitted packets. It is important to mention that the selective forwarding attack usually has most effect when the attacker is directly on the path of flowing data. But adversary can hear the neighbor packets from long distance [5].

2.3. Sinkhole attacks

The main goal of attacker in sinkhole is to attract large fractions of traffic to a region and constructing a sinkhole that the adversary is located in the center of it. For achieving this aim, Sinkhole attacks usually perform by making attractive a vulnerable node specifically to encircling nodes according to the routing algorithm. For example, an attacker can broadcast an advertisement or spoof for a very high quality route to a sink. Some protocols may try to confirm the truth of the quality of route with end-to-end acknowledgements including reliability or latency information.

The special communication pattern between sensors is one of the important reasons that sensor nodes are endangered from sinkhole attack. Since all packets in the network use and share only one base station, it is enough that compromised nodes find a single high quality route to the base station in order to influence a potentially large number of nodes [6,7].

2.4. Sybil attacks

The base of Sybil attack is that attacker can forge identities of nodes. a major side effect of this attack is to reduce the effectiveness of fault-tolerant schemes such as distributed storage, multipath routing, and topology maintenance.

Sybil attacks also pose a significant threat to geographic routing protocols. In geographical routing protocol, each node requires to transmit packet with its neighbors. So a node must have just a single set of coordinates from each of its neighbors and save them in its table but by utilizing the Sybil attack an attacker can be located in more than one situation at one time [5, 6].

2.5. Wormholes attack

In wormhole attack, attackers try to create a message appears that points away from the network. Wormhole attacks usually contain two malicious nodes that situated distant from each other. So, it can simply convince these two Separated nodes that they are neighbors by sending packets between the two of them. On the other hand, an adversary by using this attack could convince nodes that they are normally situated multiple hops from a base station that they are only one or two hops away. If an attacker is located near of sink or base station, it can interrupt routing by making a well-placed wormhole completely [6].

2.6. HELLO flood attack

Hello packets are a specific packet that usually used in many wireless sensor protocols. So, in these protocols each node needs to transmit HELLO packet for aware its neighbors, so that, when a node sends this packet, it may imagine that is located within radio range of the sender. Sometimes this assumption may be wrong. If an adversary transmits information with a sufficient power, every node in the network could be convinced that the attacker is its neighbor. This attack also can effect on protocols that based on localized information exchange between adjacent nodes like geographical routing protocol.

It is not essential for attackers to build lawful traffic due to utilize the HELLO flood attack. They can easily retransmit powerful overhead packets that every node in the network can received them [6].

2.7. Acknowledgment spoofing attack

The acknowledgment spoofing attack is designed based on this goal that a sender believes a frail connection is strong or that an unusable node is working. While nodes broadcast packet from weak or dead link, the packet may be lost. So, an attacker can prepare a selective forwarding attack utilizing acknowledgement spoofing by inspiring the certain node to send packets on those links [6,7].

3. Trust Issues

Trust and security are two important concepts that they are tightly interdependent. For example, cryptography is a modern technique for secure system that is dependent directly to a trusted key. One of the first definitions for trusted is based on Mayer, Davis and Schoorman (1995) "*the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other party will perform a particular action important to the trustor, irrespective of the ability to monitor or control the party*" [8]. In wired networks, Trust is usually provided by applying indirect trust mechanisms, such as trusted certification agencies and authentication servers. But Trust establishment in wireless sensor networks is still an open and challenging field, because these trust relationships in such a networks are extremely susceptible to attacks. Also, the absence of fixed trust infrastructure, limited resources, ephemeral connectivity and availability, shared wireless medium and physical vulnerability, make trust establishment virtually impossible. To overcome these problems, to establish trust in wireless networks should be used a number of assumptions including pre-configuration of nodes with secret keys, or presence of an omnipresent central trust authority.

Asad Amir Pirzada and et al (2004) suggested and implemented a trusted model based on an effort/return mechanism. In this model, the trust is computed based on the information that each one node can gather from the other nodes in passive mode. By analyzing the received, forwarded and overheard packets, vital information about other nodes can be collected. Possible events that can be recorded in passive mode are included Frames received, Data packets forwarded, Control packets forwarded, Data packets received, Data forwarded, Data received, etc. Information that is retrieved from these events can be grouped into one trusted category and used to compute trust in other nodes in specific situations [9].

4. Overview of protocols

According to the large number of nodes that is deployed in the many of applications of sensor networks, it is impossible to dedicate comprehensive identifiers to each node. So, it is difficult to find out the unique way among sensors that deployed randomly for transmitting data. On the other hand, non-use of specific algorithms is not definitely useful regarding energy efficiency. Routing protocols is the best method to select a group of sensor nodes and applying data collection throughout the retransmission of data has been considered [10]. One of the important routing protocols in wireless sensor networks is Geographical routing protocol. The main strategy that used in geographical routing is named greedy forwarding in which the sender transmit packet to its neighbor that is located closest to the destination. There are several ideas to define the means of nearest node to destination such as Euclidean distance to the destination, the deviation from the imaginary line between source to destination and etc. [10,11].

In Following, some geographical routing protocols are reviewed briefly.

4.1. GPSR- Greedy Perimeter Stateless Routings

The Greedy Perimeter Stateless Routing is one of the usually used location-based routing protocols for launching and maintaining a sensor network. This protocol practically functions in a stateless manner and has the capability for multi-path routing. In GPSR, it is supposed that all nodes identify the geographical position of destination node with which communication is wanted. This location information (i.e.) geographical position is also used to route traffic to its required destination from the source node through the shortest path. Each transmitted data packet from node consist the destination node’s identification and its geographical position similar two four-byte float numbers. Each node also frequently transmits a beacon to notify its near nodes relating to its recent geographical co-ordinates. The node positions are recorded, maintained and updated in a neighborhood table by all nodes receiving the beacon. To eliminate the overhead due to regular beacons, the node positions are carried onto forwarded data packets. GPSR supports two mechanisms for forwarding data packets: greedy forwarding and perimeter forwarding [12].

4.2. RGR-Receiver Based Forwarding for Geographical Routing Protocol.

Receiver Based Forwarding is an efficient approach for improving geographical sensor that is suggested and developed in December 2004 by Rodrigo Fonseca and et al. in Berkeley University. It is clear that in wireless sensor networks, when a message is transmit from one node to another; all the sender’s neighbours can hear that message. According this feature, the main difference of this idea with GPSR is in packet forwarding because instead of sender decides to forward packet, the receivers determine next hop of packet. Scilicet, when a sender wants to transmit messages, instead of addressed to a specific neighbour, the receivers recognize that whether they should forward a message or not. As mentioned earlier, the flooding issue occurs when one node receives data packet, spreads it to all its neighbours. To prevent a flooding issue in this protocol, just if the location of neighbour is closer to destination than the previous sender, the message should be transmitted again. The computation of distance between each node from destination is done by using its coordinates. Also, each message contains a header that some information likes the coordinates of last sender and final node. So, with comparing the distance of current node to destination and the distance ofpervious node to destination can decision about closeness to goal.

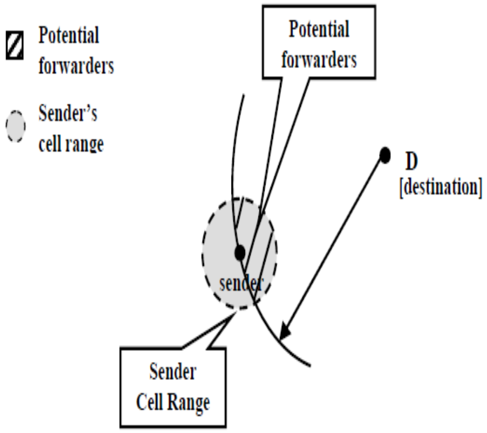


Figure1: Receiver-based protocol. Potential Forwarders (Fonseca & Sanz-Merino, 2004)

According this method, a neighbour that is nearer than sender can retransmit message but it is wasting energy. To avoid this approach, for all nodes was designed a timer which set it before forwarding message. Timer is adjusted based on closeness to destination. So, the neighbour that is closer to destination set a smaller timer than the nodes that are farther and its priority is higher. After that, if one neighbour listens to another neighbour forwarding while waiting for the timeout, it does not forward.

This routing protocol claim that it can prevent from spreading duplicates messages. So, if a node has broadcasted a copy of message, it does not retransmit following ones. To perform this, it is not sufficient to identify duplicates based on the header of message, because malicious nodes can modify the header of message. So, the identification method is carried out based on content and header of message [13].

4.3. S-GPSR–Secured Greedy Perimeter Stateless Routing Protocol.

As it mentioned before, GPSR is a routing protocol that used for geographic sensors but it is also exposed to various types of attackers. Another method which is suggested in 2010 by Samundiswary for protecting GPSR against some attacks such as sinkhole is called S-GPSR. In this method, is tried to joining trust based mechanism in the existing greedy perimeter stateless routing protocol prepares a secure routing protocol.

As mentioned in GPSR method, at first, each node during packet forwarding to a familiar destination must scan its neighbourhoods table to acquire the next hop which is optimal and leads to the goal. So, it selects the node that has the minimum distance to a specific destination. One of the newest methods for increasing the level of security in GPSR is using a trusted base approach in the neighbourhoods table to generate the most confident distance route rather than the default minimal distance. This is called S-GPSR.

The main component to implement the trust model in S-GPSR is Trust Update Interval (TUI) that used in each forward packet that is buffered in the nodes. The duration that each node should be waited before dedicating a trust or mistrust level to a node, is computed by TUI. Later than a node transmits a packet to its neighbour, it waits the neighbour's reaction for packet forwarding. So, this node faces to various situations. In the first case, the level trusted of node increased if neighbour forwards the packet in appropriate manner based on TUI. On the other hand, the level trust of node is declined if the packet is modified by the neighbour in an unsuitable way or it does not send the packet to next hop.

Each node in S-GPSR must perform two tasks, forward packet to its neighbour and control this packet. It is vital to check the integrity of forwarded packet by sender to verify the different fields in the forwarded IP packet. Therefore, confirming the acts of neighbour nods and enhancing the trust level is depended on succeeding the check of integrity. Vice versa, if the check of integrity fails or the neighbour node cannot broadcast packet, the node is treated as malicious node and the trust level decreases [14].

4.4. T-GPSR- Trusted Greedy Primeter Stateless Routing

During packet transmission to a known host, GPSR scans itsneighbourhood table to retrieve the optimal next hop leading to the destination. As there may be more than one such hopavailable, GPSR selects an adjacent neighbour that has the least distance to a particular destination. In this protocol it is attempted to modify this rule and associate the computed trust level of a node along with its geographical position in the neighborhood table due to protect GPSR against Black hole attack. In order to create the most trusted route rather than the default minimal distance route, the trust levels are utilized with the geographical distances.

To implement the trust derivation mechanism, a node buffers (GPSR Agent::buffer packet) each forwarded packet for the Trust Update Interval (TUI). The TUI is a very critical component of such a trust model and determines the time a node should wait before assigning a trust or distrust level to a no debased upon the results of a particular event. After transmission, each node promiscuously listens for the neighbouring node to forward the packet. If the neighbour forwards the packet in the proper manner (correct modification if required) within the TUI, its corresponding trust level is incremented. However, if then eighbouring node modifies the packet in a unexpected manner or does not forward the packet at all, its trust level is decremented [15].

4.5. BSR-RRS and BSR-ANS –Boundary State Routing.

Contention BSR is implemented using the combination of Greedy Bounded Compass forwarding and the Boundary Mapping. In BSR protocol, Failure of geographic forwarding due to local minima only arises on void boundaries and the outer boundary. Previous research by Karp investigated the probing of boundaries to accumulate the link state information in boundary nodes. Boundary State Routing (BSR) relies upon Greedy-Bounded-Compass forwarding. Compass forwarding selects the neighbour on the closest angle to the destination. This protocol like GPSR does not have any security feature. In order to prevent wormhole attack against BSR, two methods is designed that called BSR-RRS and BSR-ANS. Reverse Route Scheme (RRS) use hop-count technique to find malicious nodes but Authentication of Nodes Scheme (ANS) is based on authentication to find the not honest nodes [16].

5. Simulation and analysis

Simulation is one of the important steps in any survey because it allows to investigator for simulating and testing its idea in the virtual area that likes a real world. In order to simulate these routing protocols, NS-2 is selected.It is assumed that among 50 to 200 nodes are deployed randomly in 500*500 areas. In the following table some of the important parameters are mentioned.

Simulation Parameters	Values
Number of Nodes	50 and 200
Geographical environment	500*500
Size of Packet (bytes)	512
Traffic Type	CBR
Number of malicious nodes	Depend on type of attack(2-25)
Mobility model	Depend on type of routing (Static or Random way point)
Pause time(s)	20
Simulation time(s)	100

Table 1: simulation parameters

Deliver ratio is one of the useful measurement parameters in order to prove the efficiency of these secure protocol in which it is tried to calculate the numbers of packets received by destination nodes divide to the number of packets are sent by source nodes.

5.1. T-GPSR and GPSR against Black hole Attack

As it is mentioned, T-GPSR is a protocol that is designed to protect GPSR against Black hole attack. In order to check the efficiency of this protocol, 50 nodes are deployed in the area randomly. This protocol support random mobility in any way. As it is shown in the following chart, the T-GPSR has a better reaction against this attack when the numbers of malicious nodes increase. This method can improve the delivery ratio rate to 80 percent.

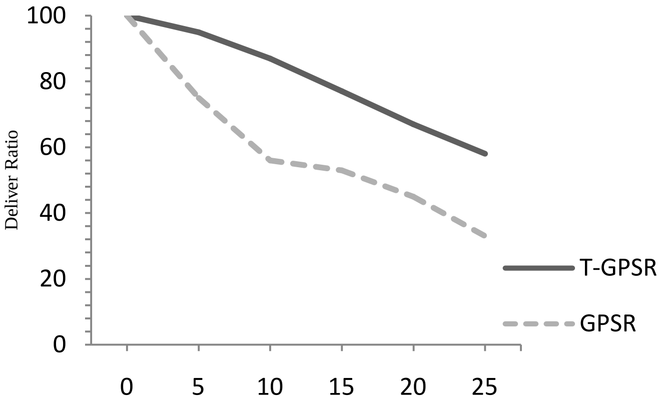


Figure 2: Deliver ratio for T-GPSR and GPSR against Black hole Attack

5.2. S-GPSR and GPSR against Selective forwarding Attack

The main purpose to design S-GPSR is to protect GPSR against selective forwarding attack by using trusted model. To evaluate this method, 100 nodes are deployed in the 500*500 (m2) environments. When the number of malicious nodes is among 5 to 15, deliver ratio of S-GPSR is about 70 percent. Finally, the rate of delivery in S-GPSR is more than GPSR, clearly.

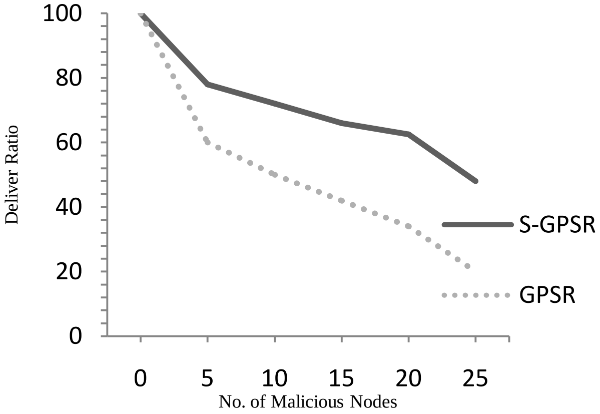


Figure3: Deliver ratio for S-GPSR and GPSR against Selective Forwarding Attack (Mobile)

5.3. RGR and GPSR against Selective Forwarding and Wormhole Attack

RGR is a method that was suggested for static geographic routing protocol based on GPSR in order to increase the security level of it. The rate of packet delivery for RGR and GPSR are illustrated in figure 4. This chart is shown thatRGR by using multipath method is more protected against selective forwarding attack than GPSR. For example, when 10 malicious nodes exist in the network, RGR's delivery ratio is higher than 80 percent, approximately. (It is important to mention that RGR is usable for static sensor networks but S-GPSR support mobility, so the comparison of them is not logical.)

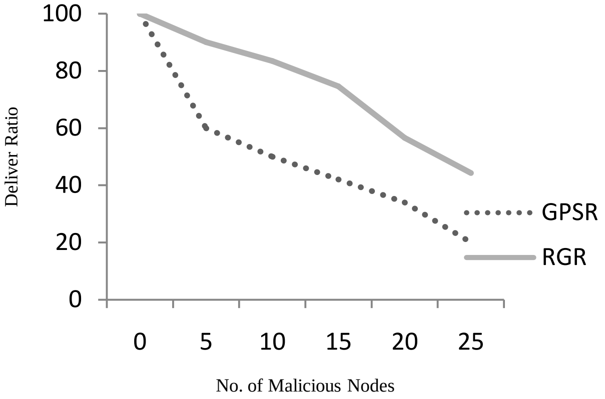


Figure 4: Deliver ratio for RGR and GPSR against Selective Forwarding Attack (Static)

In the next situation, it is tried to prove that RGR is protected against Wormhole attack but GPSR do not have any features against this attack. As it is shown in the following figure, the rate of delivery is approached to less than 10 percent when the number of malicious nodes is more than 4 in GPSR but RGR has a better reaction against this attack.

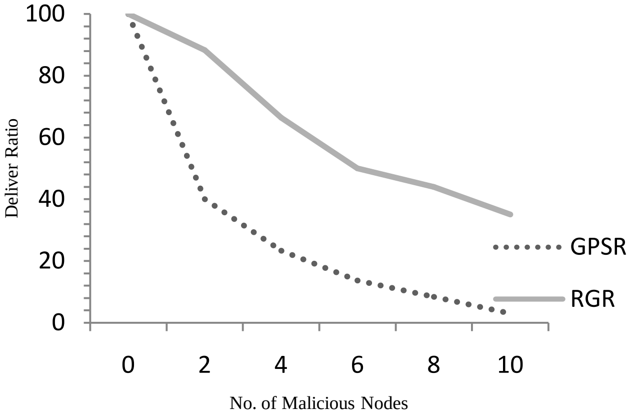


Figure 5: Deliver ratio for RGR and GPSR against Wormhole Attack (Static)

5.4. BSR, BSR-RRS and BSR-ANS against Wormhole Attack

BSR is a geographic routing protocol that used Greedy-Bounded-Compass to forward packet through destination in which there are not any security features. To secure this method against wormhole attack two method was suggested. BSR-RRS is the first method that tries to identify wormhole attack by utilizing Hop-Count technic. Based on this model, the number of hop from source to destination is compared to the number of hops through destination to source. In the next method that is called BSR-ANS, use cryptographic authentication in order to find the malicious node. The following figure is shown the rate of packet that is received in the destination to the number of packets which are sent. It is clear that ANS is the best model among these three methods by using digital signature of nodes. As it clear RRS cannot protect sensor network against wormhole attack completely.

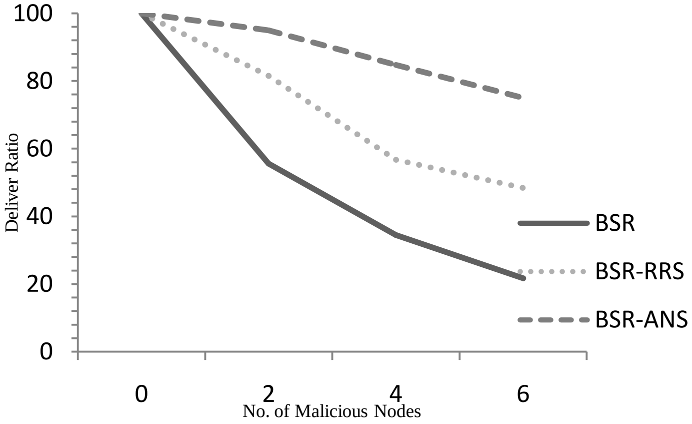


Figure 6: Deliver ratio in BSR-ANS is more than BSR-RRS and BSR against wormhole attack with mobility

6. Conclusion

In this paper, we review the secure geographic routing protocols. We also discuss about why some of routing protocols protect against some attack. Hence there are metrics to evaluate the protocols namely localization information (GPS), authentication, integrity and trust mode In order to improve their level of security. We simulated the protocols based on the delivery ratio. A qualitative comparison of secure routing protocols is summarized in table 2.

Acknowledgments

This work was supported by grant from Universiti Teknologi Malaysia,the GUP grant Vote No.Q.J130000.7182.01J52,K-economy.

Routing Protocol	Localization Information(GPS)	Authentication	Integrity	Trust model
TGPSR	✓	No	✓	✓
SGPSR	✓	No	✓	✓
RGR	✓	No	✓	No
BSR-RRS	✓	No	No	No
BSR-ANS	✓	✓	No	No

TABLE 2: QUALITATIVE COMPARISON OF SECURE ROUTING PROTOCOLS

References

[1] Palafox, L. E., & Garcia-Macias, J. (2008). Security in Wireless Sensor Networks. *IGI Global*, 547-564.

[2] Al-Karaki, J. N., & Kamal, A. E. (2004). Routing Techniques in Wireless Sensor Networks: a survey. *IEEE Communications Society, 11*, 6 - 28.

[3] Li, M., & Yang, B. (2006). A Survey on Topology issues in Wireless Sensor Network. *International Conference on Wireless Networks*, 503-510.

[4] Loo, C., Ng, M., Leckie, C., & Palaniswami, M. (2006). Intrusion Detection for Routing Attacks in Sensor Networks. *International Journal of Distributed Sensor Networks*, 313–332.

[5] Saxena, M. (2007). *SECURITY IN WIRELESS SENSOR NETWORKS - A LAYER BASED CLASSIFICATION.* West Lafayette: Center for Education and Research in Information Assurance and Security,Purdue University.

[6] Karlof, C., & Wagner, D. (2003). Secure routing in wireless sensor networks: attacks and countermeasures. *Ad Hoc Networks* , 293–315.

[7] Yick, J., Mukherjee, B., & Ghosal, D. (2008). Wireless sensor network survey. (E. Ekici, Ed.) *Computer Networks 52*, 2292-2330.

[8] Mayer, R.C. Davis, J.H. & Schoorman, F.D. (1995) An integrative model of organizational trust. Academy of management review. 20 (3), p709-734

[9] Pirzada, A. A., & McDonald, C. (2004). Establishing Trust In Pure Ad-hoc Networks. *the 27th Australasian Computer Science.26*, pp. 51-60. Dunedin: University of Otago.

[10] Akkaya, K., & Younis, M. (2005). A survey on routing protocols for wireless sensor networks. *Ad Hoc Networks 3* , 325-349.

[11] Sohraby, K., Minoli, D., & Znati, T. (2007). *WIRELESS SENSOR NETWORKS (Technology, Protocols, and Applications).* New Jersey, Hoboken, Canada: WILEYINTERSCIENCE(John Wiley & Sons, Inc.).

[12] Karp, B., & Kung, H. T. (2000). GPSR: Greedy Perimeter Stateless Routing for Wireless Networks. *6th Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom 2000)* (p. 12). Boston: MobiCom .

[13] Fonseca, R., & Merino, A. S. (2004). *Receiver Based Forwarding: Improving the security of Geographic Routing in Wireless Sensor Networks.* Berkeley: Berkeley University.

[14] Samundiswary, P., Sathian, D., & Dananjayan, P. (2010). SECURED GREEDY PERIMETER STATELESS ROUTING FOR WIRELESS SENSOR NETWORKS. *International Journal of Ad hoc, Sensor & Ubiquitous Computing(IJASUC), 1-1*, 9-20.

[15] Pirzada, A., & McDonald, C. (2007). Trusted Greedy Perimeter Stateless Routing. *ICON 2007. 15th IEEE International Conference on Networks*, 206-211.

[16] Poornima, E., & Bindhu, C. (2010). Prevention of WormholeAttacks in Geographic Routing Protocol. *International Journal of Computer Network and Security (IJCNS)*, 42-50.

Mehdi Sookhak received the B.Sc. degree in computer engineering from Azad University, Shiraz, Iran, in 2001. Now, he is student of the M.A.Sc. degree in computer Science (information security) from the University Technology Malaysia. His first publication was published in 3rd International on Engineering, Science and Humanities in Malaysia. The next one is accepted in 3rd International Conference on Computer Technology and Development, Chengdu, China, 2011. His current research interests include security of wireless sensor networks.

RaminKarimi is currently a Ph.D candidate in Department of Computer Science and Information Technology at Universiti Teknologi Malaysia, Johor, Malaysia. He received M.Sc degree in computer engineering from Iran University of Science and Technology in 2006. His research interests include Vehicular Ad Hoc Networks, Mobile ad-hoc networks, security and communication Networks.

NorafidaIthnin is a senior lecturer at Universiti Teknologi Malaysia. She received her B.Sc degree in computer science from Universiti Teknologi Malaysia in 1995, her MSc degree in Information Teknologi from Universiti Kebangsaan Malaysia in 1998 and her PHD degree in computer science from UMIST, Manchester in 2004. Her primary research interests are in security, networks, Mobile ad-hoc networks, Vehicular Ad Hoc Networks.

Mahboobeh Haghparast received M.A. degree in library and information science from Isfahan University of Iran in2010. Her research interest includes information science and information systems.

SWOT Analysis of Software Development Process Models

Ashish B. Sasankar[1], Dr Vinay Chavan[2]

[1] P.G. Department of Computer Science ,GHRIIT
Nagpur, Maharashtra, India

[2] Department of Computer Science,S.K.Porwal College,Kamptee
Nagpur, Maharashtra, India

Abstract

Software worth billions and trillions of dollars have gone waste in the past due to lack of proper techniques used for developing software resulting into software crisis. Historically , the processes of software development has played an important role in the software engineering. A number of life cycle models have been developed in last three decades. This paper is an attempt to Analyze the software process model using SWOT method. The objective is to identify Strength ,Weakness ,Opportunities and Threats of Waterfall, Spiral, Prototype etc.

Keywords: *SDLC,SWOT.*

1. Introduction

Software lifecycle models are representations of the sequence and interrelationship of broad phases within the software lifecycle. Their principal purpose is to provide a high-level plan for software lifecycle activities. They are therefore essentially management tools. The use of a software lifecycle model on a software project is important. Without the plan it provides, it can be difficult to effectively manage the project.

Within the field of Computer Science, a large number of software lifecycle models have been proposed. Each model has its own strengths and weaknesses, and each is more appropriate in certain project circumstances than others. It is generally recognised that no single software lifecycle model is appropriate in all circumstances. Because of this, for a particular software project, it is necessary to select a software lifecycle model that suits the project's characteristics. This is an important decision. The use of an inappropriate software lifecycle model can increase project costs and timescales and reduce software quality.

Now what a software lifecycle model is. Some definition are:

"framework of processes and activities concerned with the life cycle that may be organised into stages, which also acts as a common reference for communication and understanding" (ISO/IEC FDIS 12207:200726);

"A partitioning of the life of a product or project into phases." (CMMI-DEV36. This is the definition for a lifecycle model of any product or service. This may be software);

"software life cycle models serve as a high-level definition of the phases that occur during development. They are not aimed at providing detailed definitions but at highlighting the key activities and their interdependencies" (ISO/IEC TR 1975940);

"Lifecycle models describe the interrelationship between software development phases" (The NASA Software Safety Guidebook31);

2. Process Model/Life Cycle Variations

Professional system developers and the customers they serve share a common goal of building information systems that effectively support business process objectives. In order to ensure that cost-effective, quality systems are developed which address an organization's business needs, developers employ some kind of system development *Process Model* to direct the project's life cycle. Typical activities performed include the following:[1]

- System conceptualization
- System requirements and benefits analysis
- Project adoption and project scoping
- System design
- Specification of software requirements
- Architectural design
- Detailed design
- Unit development
- Software integration & testing

- System integration & testing
- Installation at site
- Site testing and acceptance
- Training and documentation
- Implementation
- Maintenance

Process Model/Life-Cycle Variations

While nearly all system development efforts engage in some combination of the above tasks, they can be differentiated by the *feedback* and *control methods* employed during development and the *timing of activities*. Most system development *Process Models* in use today have evolved from three primary approaches: *Ad-hoc Development*, *Waterfall Model*, and the *Iterative* process.

Ad-hoc Development

Early systems development often took place in a rather chaotic and haphazard manner, relying entirely on the skills and experience of the individual staff members performing the work. Today, many organizations still practice *Ad-hoc Development* either entirely or for a certain subset of their development (e.g. small projects).

The Software Engineering Institute at Carnegie Mellon University [2] points out that with *Ad-hoc Process Models*, "process capability is unpredictable because the software process is constantly changed or modified as the work progresses. Schedules, budgets, functionality, and product quality are generally (inconsistent). Performance depends on the capabilities of individuals and varies with their innate skills, knowledge, and motivations. There are few stable software processes in evidence, and performance can be predicted only by individual rather thanorganizational capability." [3]

Figure 1. Adhoc development

"Even in undisciplined organizations, however, some individual software projects produce excellent results. When such projects succeed, it is generally through the heroic efforts of a dedicated team, rather than through repeating the proven methods of an organization with a mature software process. In the absence of an organization-wide software process, repeating results depends entirely on having the same individuals available for the next project. Success that rests solely on the availability of specific individuals provides no basis for long-term productivity and quality improvement throughout an organization."[4]

2.1 The Waterfall Model

The *Waterfall Model* is the earliest method of structured system development. Although it has come under attack in recent years for being too rigid and unrealistic when it comes to quickly meeting customer's needs, the *Waterfall Model* is still widely used. It is attributed with providing the theoretical basis for other *Process Models*, because it most closely resembles a "generic" model for software development.

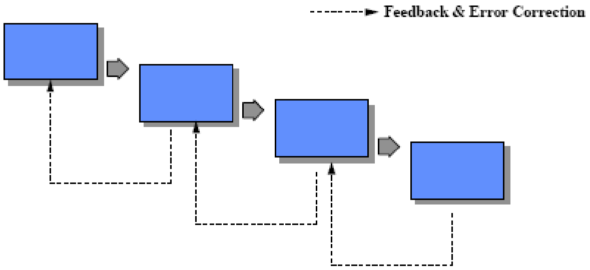


Figure 2 Waterfall model

The *Waterfall Model* consists of the following steps:

· **System Conceptualization.** System Conceptualization refers to the consideration of all aspects of the targeted business function or process, with the goals of determining how each of those aspects relates with one another, and which aspects will be incorporated into the system.

· **Systems Analysis.** This step refers to the gathering of system requirements, with the goal of determining how these requirements will be accommodated in the system. Extensive communication between the customer and the developer is essential.

· **System Design.** Once the requirements have been collected and analyzed, it is necessary to identify in detail how the system will be constructed to perform necessary tasks. More specifically, the System Design phase is focused on the data requirements (what information will be processed in the system?), the software construction (how will the application be constructed?), and the interface construction (what will the system look like? What standards will be followed?).

· **Coding.** Also known as programming, this step involves the creation of the system software. Requirements and systems specifications from the System Design step are translated into machine readable computer code.

· **Testing.** As the software is created and added to the developing system, testing is performed to ensure that it is working correctly and efficiently. Testing is generally focused on two areas: internal efficiency and external effectiveness. The goal of external effectiveness testing is to verify that the software is functioning according to system design, and that it is performing all necessary functions or sub-functions. The goal of internal testing is to make sure that the computer code is efficient, standardized, and well documented. Testing can be a labor-intensive process, due to its iterative nature.

Problems/Challenges associated with the Waterfall Model

Although the *Waterfall Model* has been used extensively over the years in the production of many quality systems, it is not without its problems. In recent years it has come under attack, due to its rigid design and inflexible procedure.

Criticisms fall into the following categories:

· Real projects rarely follow the sequential flow that the model proposes.

· At the beginning of most projects there is often a great deal of uncertainty about requirements and goals, and it is therefore difficult for customers to identify these criteria on a detailed level. The model does not accommodate this natural uncertainty very well.

· Developing a system using the *Waterfall Model* can be a long, painstaking process that does not yield a working version of the system until late in the process.

Critic

The waterfall model lacks prescribed technique of implementing management control over a project; planning, controlling, and risk management are not enveloped within the model itself. Moreover, forecasting the estimated time and cost are complicated for each stage. The life cycle can take long as the original requirements may no longer be valid, with little possibility for prototyping.

The waterfall model of system development works best when any reworking of products is kept to a minimum and the products remain unchanged. It still remains useful for steady and non-volatile types of projects, and if properly implemented, generates significant cost and timesaving. If the system is likely to go through significant changes and if the system requirements are unpredictable then different approaches are recommended, one such alternate approach is popularly know as the spiral model.

2.2 Iterative Development

The problems with the *Waterfall Model* created a demand for a new method of developing systems which could provide faster results, require less up-front information, and offer greater flexibility. With *Iterative Development*, the project is divided into small parts. This allows the development team to demonstrate results earlier on in the process and obtain valuable feedback from system users. Often, each iteration is actually a mini-*Waterfall* process with the feedback from one phase providing vital information for the design of the next phase. In a variation of this model, the software products which are produced at the end of each step (or series of steps) can go into production immediately as incremental releases.

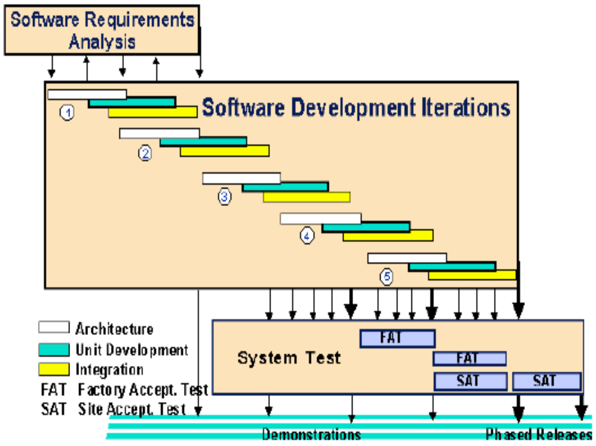


Figure 3. Iterative Development [5]

Problems/Challenges associated with the Iterative Model

While the *Iterative Model* addresses many of the problems associated with the *Waterfall Model*, it does present new challenges.

· The user community needs to be actively involved throughout the project. While this involvement is a positive for the project, it is demanding on the time of the staff and can add project delay.

· Communication and coordination skills take center stage in project development.

· Informal requests for improvement after each phase may lead to confusion -- a controlled mechanism for handling substantive requests needs to be developed.

· The *Iterative Model* can lead to "scope creep," since user feedback following each phase may lead to increased customer demands. As users see the system develop, they may realize the potential of other system capabilities which would enhance their work.

Critic

One traditional process model is the waterfall model and according to Schacchi was only accepted just until the early 1980s because of its lack of functionality. The waterfall model is said to be the easiest model to understand and I do believe with this. It is easily understood because it provides a sequential succession of phases to be followed but then it is not that reliable. Just seeing a figure of the flow of the waterfall model you would just see the sequence of phases to go through but the problem here is it would not go through a cycle but just have a one-way flow just like a waterfall. Because of its simplicity it would only be suitable for certain classes of software development and would not work well with the other software like interactive applications. This model does not have risk management and management during the life cycle and mainly document-driven or code-driven that is why it would not work as smoothly as the other model.

2.3 Variations on Iterative Development

A number of *Process Models* have evolved from the *Iterative* approach. All of these methods produce some demonstrable software product early on in the process in order to obtain valuable feedback from system users or other members of the project team. Several of these methods are described below.

Prototyping

The *Prototyping Model* was developed on the assumption that it is often difficult to know all of your requirements at the beginning of a project. Typically, users know many of the objectives that they wish to address with a system, but they do not know all the nuances of the data, nor do

they know the details of the system features and capabilities. The *Prototyping Model* allows for these conditions, and offers a development approach that yields results without first requiring all information up-front . When using the *Prototyping Model*, the developer builds a simplified version of the proposed system and presents it to the customer for consideration as part of the development process. The customer in turn provides feedback to the developer, who goes back to refine the system requirements to incorporate the additional information. Often, the prototype code is thrown away and entirely new programs are developed once requirements are identified.

There are a few different approaches that may be followed when using the *Prototyping Model*:

· creation of the major user interfaces without any substantive coding in the background in order to give the users a “feel” for what the system will look like,

· development of an abbreviated version of the system that performs a limited subset of functions; development of a paper system (depicting proposed screens, reports, relationships etc.), or · use of an existing system or system components to demonstrate some functions that will be included in the developed system.[6]

Prototyping is comprised of the following steps:

· **Requirements Definition/Collection.** Similar to the Conceptualization phase of the *Waterfall Model*, but not as comprehensive. The information collected is usually limited to a subset of the complete system requirements.

· **Design.** Once the initial layer of requirements information is collected, or new information is gathered, it is rapidly integrated into a new or existing design so that it may be folded into the prototype.

· **Prototype Creation/Modification.** The information from the design is rapidly rolled into a prototype. This may mean the creation/modification of paper information, new coding, or modifications to existing coding.

· **Assessment.** The prototype is presented to the customer for review. Comments and suggestions are collected from the customer.

· **Prototype Refinement.** Information collected from the customer is digested and the prototype is refined. The developer revises the prototype to make it more effective and efficient.

· **System Implementation.** In most cases, the system is rewritten once requirements are understood. Sometimes, the *Iterative* process eventually produces a working system that can be the cornserstone for the fully functional system.

Problems/Challenges associated with the Prototyping Model

Criticisms of the *Prototyping Model* generally fall into the following categories:

· **Prototyping can lead to false expectations.** *Prototyping* often creates a situation where the customer mistakenly believes that the system is “finished” when in fact it is not. More specifically, when using the *Prototyping Model*, the pre-implementation versions of a system are really nothing more than one-dimensional structures. The necessary, behind the-scenes work such as database normalization, documentation, testing, and reviews for efficiency have not been done. Thus the necessary underpinnings for the system are not in place.

· **Prototyping can lead to poorly designed systems.** Because the primary goal of *Prototyping* is rapid development, the design of the system can sometimes suffer because the system is built in a series of “layers” without a global consideration of the integration of all other components. While initial software development is often built to be a “throwaway, ” attempting to retroactively produce a solid system design can sometimes be problematic.

2.4 Variation of the Prototyping Model

A popular variation of the *Prototyping Model* is called Rapid Application Development (RAD).

RAD introduces strict time limits on each development phase and relies heavily on rapid application tools which allow for quick development.

Critic

Criticisms of the Prototyping Model generally fall into the following categories:

• Prototyping can lead to false expectations. Prototyping often creates a situation where the customer mistakenly believes that the system is "finished" when in fact it is not. More specifically, when using the Prototyping Model, the pre-implementation versions of a system are really nothing more than one-dimensional structures. The necessary, behindthe- scenes work such as database normalization, documentation, testing, and reviews for efficiency have not been done. Thus the necessary underpinnings for the system are not in place.

• Prototyping can lead to poorly designed systems. Because the primary goal of prototyping is rapid development, the design of the system can sometimes suffer because the system is built in a series of "layers" without a global consideration of the integration of all other components. While initial software development is often built to be a "throwaway, " attempting to retroactively produce a solid system design can sometimes be problematic.

This model cannot be used in robust application. It is convenient because it is fast from the word itself. It can replace the specification phase but not the design phase because it mainly relates to the designing phase. In the waterfall model every phase should directly right at the first time while prototyping changes frequently and the discarded if wrong.

2.5 The Exploratory Model

In some situations it is very difficult, if not impossible, to identify any of the requirements for a system at the beginning of the project. Theoretical areas such as Artificial Intelligence are candidates for using the *Exploratory Model*, because much of the research in these areas is based on guess-work, estimation, and hypothesis. In these cases, an assumption is made as to how the system might work and then rapid iterations are used to quickly incorporate suggested changes and build a usable system. A distinguishing characteristic of the *Exploratory Model* is the absence of precise specifications. Validation is based on adequacy of the end result and not on its adherence to pre-conceived requirements.

The *Exploratory Model* is extremely simple in its construction; it is composed of the following steps:

· **Initial Specification Development.** Using whatever information is immediately available, a brief System Specification is created to provide a rudimentary starting point.

· **System Construction/Modification.** A system is created and/or modified according to whatever information is available.

· **System Test.** The system is tested to see what it does, what can be learned from it, and how it may be improved.

· **System Implementation.** After many iterations of the previous two steps produce satisfactory results, the system is dubbed as "finished" and implemented.

Problems/Challenges associated with the Exploratory Model

There are numerous criticisms of the *Exploratory Model*:

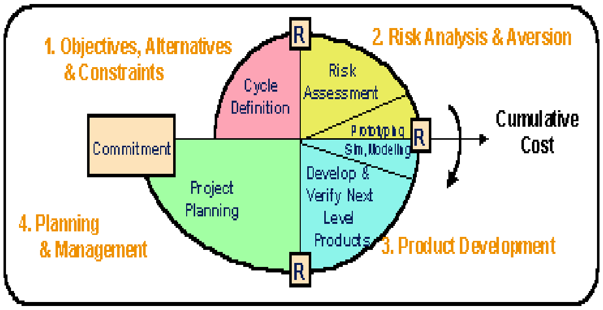
· It is limited to use with very high-level languages that allow for rapid development, such as LISP.

· It is difficult to measure or predict its cost-effectiveness.

· As with the *Prototyping Model*, the use of the *Exploratory Model* often yields inefficient or crudely designed systems, since no forethought is given as to how to produce a streamlined system.

The Spiral Model

The *Spiral Model* was designed to include the best features from the *Waterfall* and *Prototyping Models*, and introduces a new component - risk-assessment. The term "spiral" is used to describe the process that is followed as the development of the system takes place. Similar to the *Prototyping Model*, an initial version of the system is developed, and then repetitively modified based on input received from customer evaluations. Unlike the *Prototyping Model*, however, the development of each version of the system is carefully designed using the steps involved in the *Waterfall Model*. With each iteration around the spiral (beginning at the center and working outward), progressively more complete versions of the system are built.6



R=Review
Figure 4. Spiral Model[7]

Risk assessment is included as a step in the development process as a means of evaluating each version of the system to determine whether or not development should continue. If the customer decides that any identified risks are too great, the project may be halted. For example, if a substantial increase in cost or project completion time is identified during one phase of risk assessment, the customer or the developer may decide that it does not make sense to continue with the project, since the increased cost or lengthened timeframe may make continuation of the project impractical or unfeasible.

The *Spiral Model* is made up of the following steps:
· **Project Objectives.** Similar to the system conception phase of the *Waterfall Model*. Objectives are determined, possible obstacles are identified and alternative approaches are weighed.
· **Risk Assessment.** Possible alternatives are examined by the developer, and associated risks/problems are identified. Resolutions of the risks are evaluated and weighed in the consideration of project continuation. Sometimes prototyping is used to clarify needs.
· **Engineering & Production.** Detailed requirements are determined and the software piece is developed.
· **Planning and Management.** The customer is given an opportunity to analyze the results of the version created in the Engineering step and to offer feedback to the developer.

Problems/Challenges associated with the Spiral Model
Due to the relative newness of the *Spiral Model*, it is difficult to assess its strengths and weaknesses. However, the risk assessment component of the *Spiral Model* provides both developers and customers with a measuring tool that earlier *Process Model*s do not have. The measurement of risk is a feature that occurs everyday in real-life situations, but (unfortunately) not as often in the system development industry. The practical nature of this tool helps to make the *Spiral Model* a more realistic *Process Model* than some of its predecessors.

Critic
Another traditional process model is the spiral model which is suggested by Barry Boehm in 1988. Spiral model is still regarded as one of the best model because it is a combination of the prototyping model and the waterfall model and comprises the strengths of the other software models.. According to Boehm, "the major distinguishing feature of the Spiral Model is that it creates a risk-driven approach to the software process rather than a primarily document-driven or code-driven process. It incorporates many of the strengths of other models and resolves many of their difficulties" [Boehm 1988]. This model is better than the waterfall because it may allow iteration. The main concept of the spiral model is that it aims to minimize risks with the use of repeated use of prototypes so that certain changes may be applied over again if there appears a problem upon the development.

3. SWOT Analysis

3.1 Waterfall model:-

1) STRENGTH:-

- Easy adaptability by Non Technical person(End-user).
- Provides structure to inexperienced staff.
- No planning needed.
- Works well for small projects with fixed and clear requirements.
- Milestones are well defined and understood.
- Sets requirements stability.
- Good for management control (plan, staff, track).
- Works well when quality is more important than cost or schedule.
- Each phase has well defined inputs and outputs.

2) WEAKNESS:-

- All requirements must be known upfront.
- Deliverables created for each phase are considered frozen inhibits flexibility.
- Longest tangible delivery time. The customer does not see anything but the whole product when it’s ready.
- It can give a false impression of progress.
- Does not reflect problem-solving nature of software development. i.e iterations of phases.
- Integration is one big bang at the end.
- Little opportunity for customer to preview the system.
- Unsuitable for large projects and where requirements are not clear.

3) OPPORTUNITIES:-

- Requirements are very well known.
- Product definition is stable.
- Technology is understood.
- New version of an existing product.
- Porting an existing product to a new platform.
- Helpful for developing similar type of software.

4) THREATS:-

The problem with the waterfall model is that it has become hardwired into the thinking of project planners. It has become so pervasive that the requirements, design, build, and test progression is a given in most projects.
In the early days of simple, stand-alone applications, the waterfall model worked well spawning a host of voluminous methodologies, but it does not suit the

problems of the complex, risky, and integrated projects that IT has to deliver today.

IT developed stand-alone, batch applications. The complexities of integrating applications were only dreamed of by ambitious database architects. Today, hardly any development is made in isolation unless, like the NHS IT project, you give yourself the luxury of a scorched earth IT strategy. Because of its origins, the waterfall method does not address integration but ignores it until the end of the project, when we encounter the familiar task of trying to stitch together disparate applications and change schedules to the annoyance of the operations manager.

Another change in the nature of IT projects is that most of today's projects have a high proportion of reuse - implementing packages and reusing frameworks. The waterfall idea of creating a detailed set of requirements and then trying to find a package that fits is neither economic not practical. Increasingly, organisations are seeing the benefits of solution-constrained development rather than greenfield design.

The steps in waterfall model are fixed and the steps cannot change them. Model is self restricted.

If the model is not perfect, there must be some potential risks. Just as some poor descriptions and requirement changing are principal sources of project risk. In waterfall model if there is a misunderstanding in the analysis phase and that could not be found. The result could be destructive. This is almost the slowest step of development.

”The most difficult part is the communication between humans.” (Yacov, 2002).

How to manage the risks in the Waterfall model?

- It cannot be possible to avoid all the risks in the waterfall model because of the waterfall model itself. But there are still some ways to settle the problems. If team have experienced members in every job and cannot have any mistakes from the very beginning to the very end, then waterfall model is successful .
- The general method is getting prepared before the project really started. Have a essential Risk Analysis in the pre-phase can avoid the failure of every steps and rework which rise up the cost of the project.
- Making a Scheme of risk team can take a fast react in case there are some risk happened.
- Avoid the deal with the risk in surprise and make some bigger damage. Try to control every step in waterfall model.
- Do not forget to sign a contract after confirm the requirement with enduser. So that they will not ask you to add more extra functions in the software.
- Do remember that confirm there is not any mistakes and potential risks in one step. And then start your next step.
- The Project manager must take the most important point of the project. Concentrate resources on this point.
- Change the way of work from passive to active.

3.2 V-Shaped (Modified Waterfall) model:-

1) STRENGTH:-

- Emphasize planning for verification and validation of the product in early stages of product development.
- Each deliverable must be testable.
- Higher chances of success as test planning starts early in the SDLC cycle.
- Project management can track progress by milestones.
- Quickest for project where requirements are fixed and clearly defined.
- Easy to use

2) WEAKNESS:-

- Does not easily handle concurrent events.
- Does not handle iterations or phases.
- No early prototypes are available.
- Needs ample skilled resources.
- Does not easily handle dynamic changes in Requirements.
- Does not contain risk analysis activities.

3) OPPORTUNITIES:-

- Excellent choice for systems requiring high reliability.
- All requirements are known up-front.
- When it can be modified to handle changing requirements beyond analysis phase.
- Solution and technology are known.

4) THREATS:-

- The V-Shaped model is inappropriate for complex projects.

- The V-shaped model should have risk to used for large scale projects where requirements are unclearly defined and unfixed.
- The V-Shaped model should be chosen when ample technical resources are available with needed technical expertise. Since, no prototypes are produced, there is a very high risk involved in meeting customer expectations, therefore, confidence of customer should be very high in order for choosing the V-Shaped model approach.

3.3 Evolutionary Prototype model:-

1) STRENGTH:-

- Customers can “see” the system requirements as they are being gathered.
- Gains customer’s confidence as developers and customers are in sync with each other’s expectations continuously.
- Developers learn from customers.
- Ideal for online systems where high level of human computer interaction is involved.
- A more accurate end product.
- Very flexible, as changes in requirements can be accommodated much more easily with every new review and refining.
- Unexpected requirements accommodated.
- Allows for flexible design and development.
- Steady, visible signs of progress produced.
- Interaction with the prototype stimulates awareness of additional needed functionality.
- Software built through prototyping needs minimal user training as users get trained using the prototypes on their own from the very beginning of the project.
- Integration requirements are very well understood and deployment channels are decided at a very early stage.

2) WEAKNESS:-

- Tendency to abandon structured program development for “code-and-fix” development
- Bad reputation for “quick-and-dirty” methods.
- Overall maintainability may be overlooked
- The customer may want the prototype delivered.
- Process may continue forever (scope creep).

3) OPPORTUNITIES:-

- Requirements are unstable or have to be clarified.
- As the requirements clarification stage of a waterfall model.
- Develop user interfaces.
- Short-lived demonstrations.
- New, original development.
- With the analysis and design portions of object-oriented development.

4) THREATS:-

- Prototyping often creates a situation where the customer mistakenly believes that the system is "finished" when in fact it is not. More specifically, when using the Prototyping Model, the pre-implementation versions of a system are really nothing more than one-dimensional structures. The necessary, behind-the-scenes work such as database normalization ,documentation, testing, and reviews for efficiency have not been done.
- The primary goal of Prototyping is rapid development, the design of the system can sometimes suffer because the system is built in a series of "layers" without a global consideration of the integration of all other components. While initial software development is often built to be a "throwaway, " attempting to retroactively produce a solid system design can sometimes be problematic.

3.4 Rapid Application model:-

1) STRENGTH:-

- Reduced cycle time and improved productivity with fewer people means lower costs.
- Time-box approach mitigates cost and schedule risk.
- Customer involved throughout the complete cycle minimizes risk of not achieving customer satisfaction and business needs.
- Focus moves from documentation to code (WYSIWYG).
- Uses modeling concepts to capture information about business, data, and processes.
- Increases reusability of components.
- High modularization achieves a more flexible and maintainable system.
- Quick initial reviews occur.
- Encourages customer feedback.
- Integration from very beginning solves a lot of integration issues.
- Business owners actively participate

2) WEAKNESS:-

- Accelerated development process must give quick responses to the user.
- Risk of never achieving closure.
- Hard to use with legacy systems.
- Requires a system that can be modularized.
- Developers and customers must be committed to rapid-fire activities in an abbreviated time frame.
- Depends on strong team and individual performances for identifying business requirements.
- Only system that can be modularized can be built using RAD.
- Requires highly skilled developers/designers.
- High dependency on modeling skills.
- Inapplicable to cheaper projects as cost of modeling and automated code generation is very high for cheaper budgeted projects to befit.

3) OPPORTUNITIES:

- Reasonably well-known requirements.
- User involved throughout the life cycle.
- Project can be time-boxed.
- Functionality delivered in increments.
- High performance not required.
- Low technical risks.
- System can be modularized.

4) THREATS:-

- Rapid Application Development is an iterative and incremental process, there are certain risks to using RAD. It can lead to a succession of prototypes that never results in a satisfactory end product.
- The risks in RAD as opposed to "waterfall" development are related to the fact that RAD does not rely on a single requirements analysis phase.

3.5 Incremental model:-

1) STRENGTH:-

- Develop high-risk or major functions first.
- Each release delivers an operational product.
- Customer can respond to each build.
- Uses “divide and conquer” breakdown of tasks.
- Lowers initial delivery cost.
- Initial product delivery is faster.
- Customers get important functionality early.
- Risk of changing requirements is reduced.
- More flexible than waterfall.

2) WEAKNESS:-

- Requires good planning and design.
- Requires early definition of a complete and fully functional system to allow for the definition of increments.
- Well-defined module interfaces are required (some will be developed long before others)
- Total cost of the complete system is not lower.

3) OPPORTUNITIES:-

- Risk, funding, schedule, program complexity, or need for early realization of benefits.
- Most of the requirements are known up-front but are expected to evolve over time.
- A need to get basic functionality to the market early.
- On projects which have lengthy development schedules.
- On a project with new technology.

3.6 Spiral model:-

1) STRENGTH:-

- Provides early indication of insurmountable risks, without much cost.
- Users see the system early because of rapid prototyping tools.
- Critical high-risk functions are developed first.
- The design does not have to be perfect.
- Users can be closely tied to all lifecycle steps.
- Early and frequent feedback from users.
- Cumulative costs assessed frequently.

2) WEAKNESS:-

- Time spent for evaluating risks too large for small or low-risk projects.
- Time spent planning, resetting objectives, doing risk analysis and prototyping may be excessive.
- The model is complex.
- Risk assessment expertise is required.
- Spiral may continue indefinitely.
- Developers must be reassigned during non-development phase activities.
- May be hard to define objective, verifiable milestones that indicate readiness to proceed through the next iteration.

3) OPPORTUNITIES:-

- When creation of a prototype is appropriate.
- When costs and risk evaluation is important.

- For medium to high-risk projects.
- Long-term project commitment unwise because of potential changes to economic priorities.
- Users are unsure of their needs.
- Requirements are complex.
- New product line.
- Significant changes are expected (research and exploration).

4) THREATS:-

- The risk of spiral model is the events that took place that makes the project not to achieve clients requirement or what the users want.

4. Conclusions

Selecting an SDLC model can be compared in many ways to the specification of user requirements, the more data gathered and examined, the higher the chances for successful completion of the project. Just as the specifications of user requirements are vital in the stages of design and computer system development, so can the knowledge and regulations which constitute the basis for SDLC model selection determine the success or failure of a given project.

A SWOT analysis is a tool to assess and to develop strategies to remain competitive. To sum up, selecting an appropriate SDLC model is a complex and a challenging task, which requires not only broad theoretical knowledge, but also consultation with experienced expert managers.

References

[1] Kal Toth, Intellitech Consulting Inc. and Simon Fraser University; list is partially created from Software Engineering Best Practices,1997.
[2] Information on the Software Engineering Institute can be found at http://www.sei.cmu.edu.
[3] Mark C. Paulk, Charles V. Weber, Suzanne M. Garcia, Mary Beth Chrissis, and Marilyn W. Bush, "Key Practices of the Capability Maturity Model, Version 1.1," Software Engineering Institute, February 1993, p 1.
[4] Mark C. Paulk, Bill Curtis, Mary Beth Chrissis, and Charles V. Weber, "Capability Maturity Model for Software, Version 1.1," Software Engineering Institute, February 1993, p 18.
[5] Kal Toth, Intellitech Consulting Inc. and Simon Fraser University, from lecture notes: Software Engineering Best Practices, 1997.
[6] Linda Spence, University of Sutherland, “Software Engineering,” available at http://osiris.sunderland.ac.uk/rif/linda_spence/HTML/contents.html
[7] Kal Toth, Intellitech Consulting Inc. and Simon Fraser University, from lecture notes: Software Engineering Best Practices, 1997.
[8] Frank Kand, “A Contingency Based Approach to Requirements Elicitation and Systems Development,” London School of Economics, J. Systems Software 1998; 40: pp. 3-6.
[9]Bryant, A. (2000), “Chinese Encyclopaedias and Balinese Cockfights – Lessons for Business Process Change and Knowledge Management,” In *Knowledge Engineering and Knowledge Management*,
[10]Wang, Y. (2002a), “The Real-Time Process Algebra (RTPA),” *Annals of Software Engineering 14*.

Prof. Ashish B. Sasankar had done MCA, M.Phil(Comp. Sci), M.Tech(CSE) and pursuing Phd in Software Engineering from RTM, Nagpur University(INDIA). He is having 12 years of Experience in Education field. He is currently working in GHRIIT, Nagpur(India). He had published 15 international and national papers. He is member of IEEE and CSI .

Dr. Vinay Chavan had Phd ,Msc in computer science. He is working as Professor in Computer Science Dept, S.K.Porwal College Nagpur (INDIA) .

Image Restoration Using Thresholding Techniques on Wavelet Coefficients

Rubeena Vohra[1], Akash Tayal[2]

[1] Assistant Professor, BVCOE, New Delhi

[2]Assistant Professor, IGIT, New Delhi

Abstract

Image restoration from corrupted image is a classical problem in the field of image processing. Additive random noise can easily be removed using simple threshold methods with linear and non-linear filtering techniques. De-noising of natural images corrupted by Gaussian noise using wavelet techniques is very effective because of its ability to capture the energy of a signal in few energy transform values. The wavelet de-noising scheme thresholds the wavelet coefficients arising from the standard discrete wavelet transform. In this paper, it is proposed to investigate the suitability of different wavelet bases and the decomposition levels on the performance of image de-noising algorithms in terms of peak signal -to- noise ratio.

Keywords: *Image, De-noising, Wavelet Transform*

1. Introduction

Image restoration is the removal or reduction of degradations that are incurred while the image is being obtained. Degradation comes from blurring as well as noise due to electronic and photometric sources. In addition to blurring the image is often corrupted by noise during its acquisition and transmission. For example, in the image acquisition, the performance of imaging sensors is affected by a variety of factors, such as, environmental conditions and by the quality of the sensing elements themselves. For instance, in acquiring images with a CCD camera, light levels and sensor temperature are major factors affecting the amount of noise in the resulting image. Images are also corrupted during transmission, due to interference in the channel used for transmission. The main objective of de-noising techniques for random noise removal is to suppress the noise while preserving the original image details. Statistical filters like Average filter [5], [6], Median filter [7] can be used for removing such noises but the wavelet based de-noising techniques proved better results than these filters. In general, image de-noising imposes a compromise between noise reduction and preserving significant image details. To achieve a good performance in this respect, a de-noising algorithm has to adapt to image discontinuities. It compresses essential information in a signal into relatively few, large coefficients, which represent image details at different resolution scales. In recent years there has been a fair amount of research on wavelet thresholding and threshold selection for signal and image de-noising [2] because wavelet provides an appropriate basis for separating noisy signal from image signal. Many wavelet based thresholding techniques like VisuShrink, SureShrink have proved better efficiency in image denoising. We describe here an efficient thresholding technique for denoising by analyzing the statistical parameters of the wavelet coefficients. The threshold is estimated and the coefficients are killed or remain unchanged or shrinked, depending on the type of thresholding (i.e. hard or soft). The first method estimates the threshold level by a median estimator, which implements the noise standard deviation from the coefficients of the diagonal subband of the first level (i.e. HH) and is called global. The second method refers to a median estimator, which is applied on all the detail coefficients of each level, so is level dependent. Eventually, the third approach employs a median estimator which is applied on the horizontal-vertical-diagonal detail coefficients of each subband, so is detail dependent. This paper is organized as follows. Section 2 is a brief review of the discrete wavelet transform. In section 3, the concept of wavelet thresholding is developed. Section 4 explains the proposed method of de-noising based on wavelet decomposition. Experimental evaluation is performed in section 5 and finally conclusions are given in section 6.

2. Discrete Wavelet Transform (DWT)

The mathematical approach to the discrete wavelet transform (DWT) is based on the fact that a function f (t) can be linearly represented as:

$$f(t) = \sum_k a_k \psi_k(t) \qquad (1)$$

where a_k are the analysis coefficients and ψ_k the analyzing functions, which are called basis functions, if the above analysis is unique. If the basis functions are orthogonal, that is,

$$\langle \psi_k(t), \psi_l(t) \rangle = \int \psi_k(t)\psi_l(t)dt = 0 \quad \text{for } k \neq l \qquad (2)$$

the coefficients can be estimated from the following equation:

$$a_k = \langle f(t), \psi_k(t) \rangle = \int f(t)\psi_k(t)dt \qquad (3)$$

where *f(t)* is given from (1). In general, a 2-D signal may be transformed by DWT as:

$$f(t) = \sum_k \sum_j \alpha_{j,k}\psi_{j,k}(t) \qquad (4)$$

where $\alpha_{j,k}$ and $\psi_{j,k}$ are the transform coefficients and basis functions respectively. Equation (4) is the inverse transform, given by $\alpha_{j,k}$ and $\psi_{j,k}$. Therefore, a function f(t) may be represented by transform coefficients, which are estimated from the internal product of that function with an orthogonal basis function. Inversely, the desired function may be reconstructed from these coefficients and the basis function. These basis functions are called wavelets [1], [3].

Another consideration of the wavelets is the subband coding theory or multiresolution analysis [4]. The first component to multiresolution analysis is vector spaces. For each vector space, there is another vector space of higher resolution until you get to the final image. The basis of each of these vector spaces is the scale function for the wavelet. We can consider an image a vector space such as V_j would be perfectly normal image and V_{j-1} would be that image at a lower resolution until we get V_o where there is only one pixel in the entire image. For such vector space V_j there is an orthogonal compliment called W_j and the basis function for this vector space is the wavelet. If the function $\phi(x) \in Vo$ such that the set of functions of $\phi(x)$ and its integer translates $\{\phi(x-k)/k \in z\}$ forms a basis for space V_o which is termed as scaling function or father function. The subspace Vj are nested which implies $V_j \in V_{j+1}$. It is possible to decompose V_{j+1} in V_j and W_j.

$$V_j \oplus W_j = V_{j+1} \qquad (5)$$

Also, $W_j \in V_j$ (6)

Ψ (x) $\in$ Wo obeys translation property such that Ψ(x-k) $\in$ Wo, k $\in$ z [11] . form a basis function for space W_o which is termed as wavelet function or mother function. DWT scaling function for 2-D DWT can be obtained by multiplying two 1-D scaling functions: $\phi(x,y) = \phi(x)\phi(y)$. Wavelet function for 2-D DWT can be obtained by multiplying two wavelet functions. For 2-D case there exists three wavelet functions that scan details in horizontal Ψ (x, y) = Ψ (x) $\phi(y)$, vertical Ψ (x, y) = $\phi(x)$ Ψ (y) and diagonal direction Ψ (x, y) = Ψ (x) Ψ (y). As a result, there are three types of detailed images for such resolution: horizontal, vertical and diagonal.

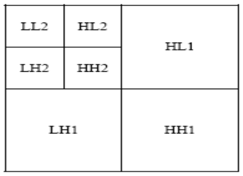


Fig 1: Two- level decomposition

3. Wavelet Thresholding

Let f= {fij, i, j=1, 2 ...M} denotes a M x M matrix of original image to be recovered and M is some integer power of 2. During the transmission, the signal f is corrupted by independent and identically distributed zero mean, white Gaussian noise nij with standard deviation σ i.e. nij ~ N (0, σ^2) and at the receiver end, the noisy observation gij=fij+nij is obtained. The goal is to estimate the signal f from the noisy observations gij such that the Mean Square Error (MSE) is minimum. To achieve this gij is transformed into wavelet domain, which decomposes the gij into many subbands, which separates the signal into so many frequency bands. The small coefficients in the subbands are dominated by noise, while coefficients with large absolute value carry more signal information than noise. Replacing noisy coefficients (small coefficients below certain value) by zero and an inverse wavelet transform may lead to reconstruction that has lesser noise. Normally Hard Thresholding and Soft Thresholding techniques are used for such denoising process. Hard and Soft thresholding [14] with threshold λ are defined as follows. The hard thresholding operator is defined as:

$$D(U,\lambda) = U \ \ for\ all\ |U| < \lambda \qquad (7)$$

= 0 otherwise

The soft thresholding operator on the other hand is defined as:

$$D\,(U, \lambda) = sign(U)*max(0,|U|-\lambda) \qquad (8)$$

4. De-noising Algorithm

1. Transform the noisy image into orthogonal domain by discrete 2D wavelet transform.

2. Apply hard or soft thresholding the noisy detail coefficients of the wavelet transform.

3. Perform inverse discrete wavelet transform to obtain the de-noised image.

Here, the threshold plays an important role in the de-noising process. Normally, hard thresholding and soft thresholding techniques are used for such de-noising process. Hard thresholding is a keep or kill rule whereas soft thresholding shrinks the coefficients [13] above the threshold in absolute value. It is a shrink or kill rule. The following are the methods of threshold selection for image de-noising based on wavelet transform:

4.1 Visushrink

It is the de-noising technique introduced by Donoho [12], [8] , it uses the threshold value t that is proportional to standard deviation of noise follows "Hard Thresholding Rule". The universal rule for threshold T can be calculated using the formulae,

$$T=\sigma\sqrt{2}\log n \quad (9)$$

This method performs well under a number of applications because wavelet transform has the compaction property of having only a small number of large coefficients. All the rest wavelet coefficients are very small. This algorithm offers the advantages of smoothness and adaptation. However, it exhibits visual artifacts.

4.2 Sureshrink

A threshold chooser based on Stein's Unbiased Risk Estimator (SURE) was proposed by Donoho and Johnstone and is called as SureShrink. It is a combination of the universal threshold and the SURE threshold. This method specifies a threshold value tj for each resolution level j in the wavelet transform which is referred to as level dependent thresholding. The goal of SureShrink is to minimize the mean squared error, defined as

$$\text{MSE}=1/\text{MN}\sum_{y=1}^{M}\sum_{x=1}^{N}[I(x,y)-I^{'}(x,y)]^{2} \quad (10)$$

where $I^{'}(x,y)$ is the estimate of the signal while I(x,y) is the original signal without noise. The SureShrink suppresses noise by thresholding the empirical wavelet coefficients. The SureShrink threshold t* is defined as:

$$t^{*}=\min(t,\sigma^{2}\log n) \quad (11)$$

where t denotes the value that minimizes Stein's Unbiased Risk Estimator, σ is the noise variance computed from Equation (11), and n is the size of the image. The Sureshrink method follows the soft thresholding rule. The thresholding employed here is adaptive, i.e., a threshold level is assigned to each dyadic resolution level by the principle of minimizing the Stein's Unbiased Risk Estimator for threshold estimates. It is smoothness adaptive, which means that if the unknown function contains abrupt changes or boundaries in the image, the reconstructed image also does.

4.3 Bayesshrink

BayesShrink was proposed by Chang, Yu and Vetterli [10]. The goal of this method is to minimize the Bayesian risk, and hence its name, BayesShrink. It uses soft thresholding and is subband-dependent, which means that thresholding is done at each band of resolution in the wavelet decomposition. Like the SureShrink procedure, it is smoothness adaptive. The Bayes threshold, tB, is defined:

$$tB=\sigma^{2}/\sigma_{s} \quad (12)$$

where σ^{2} is the noise variance and σ is the signal variance without noise. The noise variance σ^{2} is estimated from the subband HH1 by the median estimator. From the definition of additive noise we have

$$w(x,y)=s(x,y)+n(x,y). \quad (13)$$

Since the noise and the signal are independent of each other, it can be stated that

$$\sigma_{w}^{2}=\sigma_{s}^{2}+\sigma^{2} \quad (14)$$

σ_{W}^{2} can be computed as shown below:

$$\sigma_{w}^{2}=1/n^{2}\sum_{x,y-1}^{n}w^{2}(x,y) \quad (15)$$

The variance of the signal σ_{S}^{2} is computed as:

$$\sigma_{s}=\sqrt{\max(\sigma_{w}^{2}-\sigma^{2},0)} \quad (16)$$

5. EXPERIMENTAL RESULTS

The above said methods are evaluated using the quality measure Peak Signal to Noise ratio which is calculated using the formulae,

$$\text{PSNR}=10*\log_{10}(255)^{2}/\text{MSE(db)} \quad (17)$$

where MSE is the mean squared error between the original image and the reconstructed de-noised image. Quantitatively assessing the performance in practical application is complicated issue because the ideal image is normally unknown at the receiver end. So this paper uses the following method for experiments. One original image is applied with Gaussian noise with variance value 0.001. In this paper, different wavelet bases are used in all methods. For taking the wavelet transform of the image, readily available MATLAB routines are taken. In each sub-band, individual pixels of the image are shrinked based on the threshold selection. A de-noised wavelet transform is created by shrinking pixels. The inverse wavelet transform is the de-noised image. In this paper three images of different sizes are denoised by applying the techniques discussed above. The simulation results are shown below of three images: “imde1.jpg”,”imde2.jpg”, “imde3.jpg” with its original image, noise corrupted image with Gaussian noise at variance 0.001 and its denoised image.

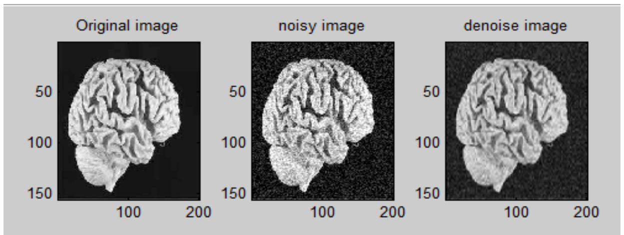


Fig.2:“imde1.jpg”with its original, noisy and denoised pattern

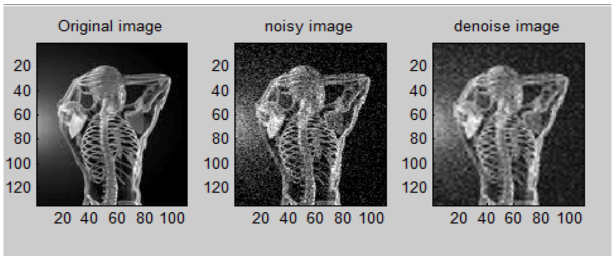


Fig.3:“imde2.jpg”with its original, noisy and denoised pattern

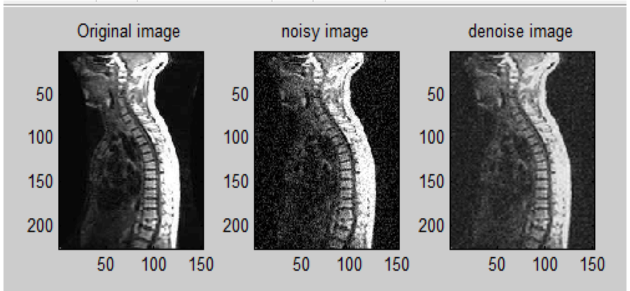


Fig.4:“imde2.jpg”with its original, noisy and denoised pattern

Table 1: signal-to-noise ratios of the thresholding techniques compared to weiner filter

Wavelet	Image	Figure1	Figure2	Figure3
	Median	31.6	35.2	33.5
haar	Visushrink	29.2	32.8	30.8
	Sureshrink	31.8	35.0	32.8
	Bayeshrink	30.0	33.3	31.4
Db16	Visushrink	30.4	33.4	31.4
	Sureshrink	32.8	36.4	33.7
	Bayeshrink	29.0	33.3	30.4
Coif5	Visushrink	30.5	33.8	31.5
	Sureshrink	**33.0**	**36.2**	**34.0**
	Bayeshrink	29.9	33.4	30.4
Sym8	Visushrink	30.5	33.5	31.6
	Sureshrink	32.9	36.0	33.8
	Bayeshrink	30.0	34.0	32.2

Throughout the text, we tried to present numerous original interpretations, pictorial explanations and discussions broadening our viewpoints on this topic. The use of the localized context-dependent hard and soft thresholding operators have resulted in some improvement in the performance of the various standard wavelet thresholding methods studied in this paper.
For the above mentioned three methods, image de-noising is performed using wavelets for the second level decomposition and the results are shown in figure1, figure2 and figure3 along with the table formulated for noise variance 0.001. Along with the comparison to the Weiner Filter “Sureshrink” gave the best possible results.

6. CONCLUSION

In this paper, the image de-noising using discrete wavelet transform is analyzed. The experiments were conducted to study the suitability of different wavelet bases and also different window sizes. Among all discrete wavelet bases, coiflet performs well in image de-noising. For Gaussian noise (0, 0.09) – PSNR improves by the use of Hard Thresholding technique. Experimental results also show that Sureshrink gives better result than Visushrink and Bayesshrink as compared to Weiner filter.

References

[1] M. Vetterli, J. Kovacevic,Wavelets and subband coding, Englewood Cliffs, NJ, Prentice Hall, 1995.

[2] C.S.Burrus, R.A.Gopinath, H.Guo, “Introduction to Wavelets and Wavelet Transforms”, Prentice Hall, 1998, pp. 2-18.

[3] G. Strang,T. Nguyen, “Wavelets and Filter Banks”, Wellesley, 1997.

[4] S. Mallat, “A theory for multiresolution signal decomposition: the wavelet representation”, IEEE Trans. PAMI, vol. 11, no. 7, pp. 674-693.

[5] Maher A. Sid-Ahmed. (1995). Image Processing-Theory algorithm and architecture. McGraw-Hill, pp 78-80.

[6] Rafael C.Gonzalez & Richard E.Wodds. (1993). Digital Image Processing. Addison Wesley publishing Company , pp 41-43.

[7] D.L. Donoho. (1994). Ideal spatial adoption by wavelet shrinkage. Biometrika, volume 81, pp.425-455.

[8] D.L. Donoho and I.M. Johnstone. (1995). Adapting to unknown smoothness via wavelet shrinkage. Journal of American Statistical Association., Vol. 90, no. 432, pp1200-1224.

[9] S. Grace Chang, Bin Yu and M. Vattereli. (2000). Wavelet Thresholding for Multiple Noisy Image Copies. IEEE Transaction. Image Processing, vol. 9, pp.1631- 1635.

[10] S. Grace Chang, Bin Yu and M. Vattereli. (2000). Spatially Adaptive Wavelet Thresholding with Context Modeling for Imaged noising. IEEE Transaction - Image Processing, volume 9, pp. 1522-1530.

[11] M. Vattereli and J. Kovacevic. (1995). Wavelets and Subband Coding. Englewood Cliffs. NJ, Prentice Hall.

[12] Maarten Janse. (2001). Noise Reduction by Wavelet Thresholding.Volume 161, Springer Verlag, United States of America, I edition.

[13] Carl Taswell. (2000). The what, how and why wavelet shrinkage denoising. Computing in science and Engineering, pp.12-19.

[14]D.L. Donoho. (1995). De-noising by soft thresholding. IEEE Transactions on Information Theory, volume 41, pp.613-627

The Regulatory Challenges and Opportunities of IP Telephony: Perspective on Malaysia

Khaled-Shukran

Asia Europe Institute, University Malaya, Kuala Lumpur
50603, Malaysia

Abstract

IP telephony is a transmission of voice and data over Packet Switched IP Networks and it has become a key issue in the telecommunication industry worldwide because of its higher efficiency and cheapest call rate. Hence, most of the countries in the world set up their telecommunication platform on IP based network as a fast revenue generating sources. Though Malaysia is on the way forward on IP based platform but the growth is not reached at satisfactory level due to the absence of effective regulatory frameworks policies. So, growth of IP telephony is declining and foreign joint investors are facing difficulties in Malaysia. Besides, local incumbent operators are gaining market advantages of having huge subscribers locally and providing service in a form of direct calling system. So, this paper tried to explore the issue of "obstacle and Survive" and reasons of declining IP telephony service providers in Malaysia.

Keywords*: IP networks, regulatory framework, incumbent operator, market advantages, IP telephony.*

1. Introduction

The development of Internet Protocol (IP) in mid -1970's has started a whirlwind of change in the telecommunications market in the world. IP development has opened a wide range of services in global communications. This service offers alternative cheapest voice call that evades Public Switched Network (PSTN). It is capable of providing higher efficiency and lower cost for communication to the consumers as well as end-users. VoIP service offers *"Everything over IP"* based platform. Nowadays, dual-mode handset and its services offer VoIP calls over wire or wireless mode. BT (Fusion) in UK, T-Mobile in Germany (At home) and Orange (Unik) in France are the best examples of this type of system. However, VoIP is going to be a mainstay in the corporate world and Japan announces that, currently 60% subscribers are using VoIP service. In 2007, there were 100 million VoIP subscribers in the world but over the time increasing very fast. It assumes that, at the end of 2011, there will be 250 million VoIP subscribers throughout the world. Following the trend of IP development, Malaysia is also sets the VoIP infrastructure where service providers are rapidly expanding their cost effective service. Malaysian Communications and Multimedia Commission (MCMC) along with local incumbent operators are facilitating to the foreign investors to commence IP telephony services. Telecom Malaysia Berhad (TM) is the leading operator in Malaysia where it contributes the higher revenue in this sector. Alongside, Redtone Sdn Bhd and Marchantrade Asia Sdn Bhd also providing IP telephony service in Malaysia. In order to start this service there is pre-requite to have an ASP license from Malaysian Communications and Multimedia Commission (MCMC) and every year has to renew. In every year, ASP licensed company is increasing very fast. At present it has crossed 400 ASP license holders' local as well as foreign companies in Malaysia.

With the forward movement of VoIP, everyone facilitates to the use of IP telephony services because of its reliability and lowest cost. Though IP telephony takes a new place in the global communication even in Malaysia incredibly, there are so many obstacles to establish IP Telephony business in Malaysia for foreign investors.

1.1 Research Objectives

The objective of this research is to focus on the difficulties to run IP telephony business in Malaysia for foreign investors or companies. Due to the regulatory and pricing challenges, foreign investors are facing problems to establish the business in Malaysia. For this, this paper is specially focuses on regulatory issues and tried to find out a way to overcome these problems to continue IP Telephony business in Malaysia. This paper also conducted research upon a number of countries to set its opinion and recommendations.

1.2 Problem Statement

VoIP is getting popular all over the world because of higher efficiency and lowest call rates. Unlike Japan, Australia, USA and others countries, Malaysian service providers are charging more [SPL high rate] to the ASP service holders or companies. Reducing the cost is one of the main issues of IT business, therefore, everybody will find out the cheapest call rate and best sound quality. Japan, USA and Australia offers low price that's why ASP service providers get benefit from them. In the competitive market place outside carriers offers low cost whereas TM offers high rate. According to MCMC regulations ASP license holders are bound to buy minutes volume either TM or any other local incumbent operators. Price could be less if the volume of usage increased but normally ASP service providers cannot increase volume usage at the starting of their business because of not having huge clients. However, in order to reduce the rate, IP telephony companies are highly suggested by TM to increase the volume of minutes. Nevertheless, service providers are not able to increase the volume usage with a high rate because of the global price differences with Malaysian incumbent operators. Another issue is toll free-number which is used for calling card and call back services. Unlike Japan, in Malaysia toll-free is allowed only from fixed phone numbers. But if there is an opportunity to use toll-free through hand phone, public phone, any booth or university booth foreign companies can get at least marginal profit from their service.

So due to the enormous encounters and not having sufficient facilities, IP telephony service providers are declining from the market or switched to other services, or even totally washout from the marketplace. In this regard, this paper tried to find the ways to overcome all these problems.

1.3 Scope of Research

At present, Malaysia is enhancing their capabilities to create unique opportunities to be a part of the global information society. ICT development, innovation, economic incentive and information structure are the main factors for advancement. Moreover, in the expansion of the ICT, competition appears in the global market and countries are shifting from old paradigm like physical and monetary asset to the new information age. So ICT along with telecommunication sector opens up the new horizon of opportunities and attract foreign companies to invest in this country. Malaysia invites foreign investors in ICT sector by providing secure place which will be the rapid revenue generating source towards the development process. If, foreign companies come to set up their IP telephony business, massive employment opportunities will be created and country will be achieving a developed nation status.

1.4 Research Questions

- What are the obstacles to run IP telephony business for foreign companies?
- To what extend Telekom Malaysia Berhad can take initiative to solve the problem as TM plays an important role regarding IP telephony business?
- Should new rules for ASP license holders adopt by MCMC to give more facilities to the foreign as well as local companies?
- The topic raise the important questions "obstacle and survive" and what are the necessary steps need to measure MCMC to make it easier?

1.5 Significance of the Research

The tremendous use of Internet Protocol (IP) networks for communication service, especially, in telephony become essential part for the telecommunications industry worldwide. The major key issue appears into spotlight for ICT policy maker and regulatory authority that, IP based network communications reduced the cost thus it becomes the technology of choice.

So the growths of IP telephony networks around the world are becoming popular and bring broaden implication in the telecommunication industry. Consequently, the major international Public Telecommunications Operator (PTOs) has taken initiatives to migrate all their international traffic onto IP platforms and made a substantial investment in this sector. It is evident that, IP based network will reduce quarter of cost than circuit switched networks. So the issue of cost reducing and less infrastructure cost are the predominant part in IP telephony sector. Malaysian government has taken initiative on the way forward to the IP based platform as it is highly increasing and lower cost of IP networks for communications.

Though Internet protocol (IP) telephony is rapidly reaching at the top of the agenda in the telecommunication industry worldwide, there are relentless problems are facing ASP license holder companies, such as; pricing, volume usage, toll free from mobile, direct competition with local incumbent operator and others. So, in Malaysia, this research will help to create the awareness of the Malaysian Communications and Multimedia Commission to take initiatives for the ASP licensed companies so that, they are able to run their business effectively and initiate to protect them from declining or scrub down from the telecommunication market and contribute to continuous economic development in Malaysia. This research will help to think to the government about VOIP service providers in Malaysia which will brings potential benefit for the country. This Research would be the supporting tool for adapting new laws in favor of foreign IP telephony companies which will increase the usage of IP based networks in the long run and will help ASP license holders companies enhancing their business industry in IP telephony networks. It will also help foreign people/investors to increase overall performance. Finally this research would change the attitude towards IP telephony companies which will have a positive effect towards its adaptability and manage VoIP companies effectively.

2. Background Study and Fundamental Changes of IP Telephony

Internet Telephony or Voice over Internet Protocol landscape is rapidly changing in the telecommunication industry. It is a fundamental archetypal shift from traditional PSTN (Public Switched Telephone Network) Circuit –Switched Voice Networks to Packet Switched Data Network where Internet protocol is predominant. After the invention of IP telephony, its popularity in the world of voice communication is fast and reliable. It meets the higher performance with lower voice networking cost. As revenue generating business source IP telephony can meet the service providers as well as user's expectation. So IP telephony is able to prove its capabilities of delivering service to the end users and increase business opportunities in the world telephony market because of its lower voice, equipment and administrative cost. So over the time IP telephony is becoming the driving force of IP based communications. According to the Malaysian IP Telephony Industry Report 2007,

"*Operators of IP telephony are 'widening' over time – from the circle of pure VOIP players who do not own a network and offers a voice call service to the larger arena of incumbent operator offering broadband access over which IP telephony can be offered free or at lower charges in combined packages''* [1].

However, IP telephony is putting forward of innumerous possibilities of the fixed line operators towards the new revenue tributary of telecommunication industry. As fixed line revenue landscape is declining and IP telephony networks are generating more revenue that's why alliance with service providers and content providers are increasing. Consequently, it becomes the stronghold of the commercial as well as business world through IP telephony business enterprise. Thus, IP telephony appears as a continual and prospective threat for the fixed line operators. So, in order to exist in the world market, migration towards next generation network followed by IP based network technology is absolutely desirable.

To meet the expectation of customer demand and reliability, network infrastructure must provide high performance with low cost. IP telephony is providing high quality of voice transfer using low bandwidth. Therefore it gives high priority of usage affirmation. As a result end-users reliability is increasing to meet expectations of clients.

In addition, regulatory issues of IP telephony create problems for the service provider in certain countries in the world where some are facing the banning to provide services. Though there is no banning or license restrictions in Malaysia on providing services but few regulatory issues need to acclimatize favoring to Application Service Providers, so that they can continue their services without any difficulties. In effect of regulatory issues along with TM high pricing matter for call termination and provide same service in identical destination especially in South Asian Countries in a form of direct calling with promotional offers, ASP holders faces threat and plunge in a disastrous situation to continue their IP telephony services in Malaysia.

Even the new market of IP telephony is not expanding due to the extreme competitive attitude of incumbent operators. Usually competitive phenomenon helps to enhance the market growth but, if competition appears through local incumbent operators then ASP holders will face more difficulties.

Moreover, few issues are highly notified in terms of VoIP enterprises, such as, problems in network and security factors especially for callback and callshop solutions. A report stated that, about 27% of Asia enterprises are worrying about the security issue because of using public networks where packets pass through by any router and anyone can access easily [2]. So, in terms of call back and call shop solutions it is the main concern to provide services. Service providers may face severe hacking problem and loss huge amount of purchased call volume due to the security issue.

The report again focuses on the converged IP network that, though 55% of Asian enterprises have embraced this network, it is critical for the networking environment. It assumes that, converged voice, data along with video will be the attractive features of future telecommunications industry.

In order to build a good business environment, call availability in VOIP services is mandatory. Besides, consumer protections, security along with universal service provision are also important for building the VoIP infrastructure. Giving the broad idea of IP telephony or

Voice over Internet Protocol this chapter addressed the overview of IP telephony, IP Telephony architecture, fundamental change in the communication, the growth of IP telephony and its development over the time, IP telephony growth in Malaysia, revenue trends, business value, issue and challenges, market drivers in IP telephony, fundamental process and finally business and economic importance of IP telephony.

2.1 IP Telephony Overview

The term "IP Telephony" can be defined in different way as there is no exact definition of it. It is a service for international call or international bypass or it is a network for the '*next generation signaling and multimedia connectivity'* [3]. IP telephony includes a set of technologies that allows data to transmit and collaborate with IP based network such as LANs (Local Area Network), WANs (Wide Area Network and Metro Area Network) where broadband connection is prerequisite. This term also been used interchangeably with VoIP (Voice over Internet Protocol) which offers cost effective technique of communication. According to William A.Yarberry,Jr.**;** 'IP telephony is a (i) set of standards for packet transmission;(ii) ability to commingle various media such as voice, data, and video, on LANs, WANs and the internet; and (iii) the flexibility with regard to physical media-IP telephony works over twisted pair, fiber, xDSL, ISDN, leased lines, coaxial cable and others [4]. Moreover, IP telephony uses IETF platform (The Internet Engineering Task Force) whose main concern is to provide smooth operation work of the internet in order to make internet exertion better by ensuring the high quality of service, and ITU (The International Telecommunication Union) whose main intention is to improve the telecommunication infrastructure and establish high standard of service. Hence, in 2001 ITU make a distinction of IP telephony and VoIP. IP telephony is a *"voice over IP based networks irrespective of ownership' and VoIP is a –'Voice service over networks competing with incumbent operator''* [5].

Although ITU distinguishes between the two, IP telephony is also known as VoIP, mostly use for communication because of its cost saving, flexibility and lower management cost. Vendors or service providers are interested in VoIP business because of its less financial investment and cost effective services. It also provides the 'wider and diverse range of multimedia services and innovative applications and particularly to be able to compete effectively in future E-Commerce markets' [6]. ITU defined IP telephony in such a way that;

''The Internet and IP based networks are increasingly being used as alternatives to the public switched telephone network. Internet Telephony service providers (ITSPs) can provide voice and fax services which are close to becoming functionally equivalent to those provided by public telecommunication operators (PTOs). However, few ITSPs are licensed by national authorities and they generally do not have any universal service obligations. Many countries ban IP telephony completely, yet IP calls can be made to almost any telephone in the world. Many PTOs are establishing their own IP telephony services, and/or using IP-based networks as alternative transmission platforms. In the longer term, as more and more voice traffic becomes IP data traffic, there will be little to distinguish between IP telephony and circuit – switched telephony. However, many telecommunications regulatory schemes depend upon such a distinction, both physically and as a matter of policy and law. As these trends continue, the telecommunication framework will come under increasing pressure to adapt'' [7]

There is an assumption that, the whole world will be turning to the IP telephony platform due to its innovativeness, dynamism and the great source of revenue generating phenomenon.

The anticipation of ITU is absolutely true about IP telephony. Japan reported that, about 60% subscriber using VoIP for personal communication as well as corporate use and its best example is Skype. Presently, VoIP is enormously using in Small and Medium Size Enterprises (SMEs) to enhance them. However, IP telephony is using extensively in IP-PBX, unified communication, contract centers and carrier services. Packet Switched Connection from the internet to exchange voice, fax and data transfer as an alternative of PSTN is called dedicated circuit switched connections.

IP telephony or VoIP also called peer to peer VoIP which '*gets the strengths from each individual node, adding bandwidth and processing power with each new member for the good of the many'* [8]. Garrie & Rebecca Wong (2009) defined VoIP, as a "conveyance of voice, fax and unrelated services publicly or wholly over packet switched IP-based networks including peer to peer VOIP and VOIP service connected to PSTN"[9]. It also refers to the telephony application 'that are enabled in a homogeneous IP environment as well as the integration of these applications with mainstream business process' [10]. Sangoma Technologies Corporation a Canada based company is providing voice and data connectivity components for software based communication applicant.

Though variety of new technology in telecommunications begins with a wide range of commercial services in a form of IP telephony or VoIP, regulatory issues has been grappled up with a broad implications in VoIP market.

2.2 IP Telephony Architecture

Looking at the IP telephony architecture, there are few important features are significant as William A. Yarberry, Jr. Stated;

1. Bandwidth is used more efficiently
2. Components are more evenly distributed
3. Growth is incremental
4. Port limitations are reduced
5. The architecture are less proprietary
6. Fewer single points of failure
7. Potential for lower cost
8. Single wiring

In IP, telephony packet based voice transmission can be used efficiently and at the same time number of conversation can pass through dedicatedly. Even bandwidth requirements are possibly can be reduced while the voice and data pass through. IP telephony architecture is linked with the local area network where servers are distributed in different segments. Besides, the growth of IP telephony is incremental.

In traditional PBX system users need to add additional hardware whereas in IP telephony, extension of the switch is needed and it is connected with Ethernet followed by RJ45 connectors. If it is connected to any types of devices of IP telephony port limitation issues can be reduced.

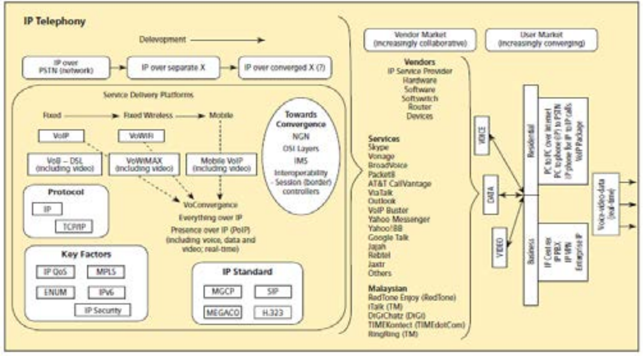
But IP telephony with TDM architecture port limitation can solve by upgrading the system. In Traditional PBX architecture is proprietary based which is complex but the IP telephony architecture less proprietary and vendors are being pressured to its standard services. TDM based PBX, backup system is very expensive and it always involves in advance cross wiring but in IP telephony architecture servers and switches can be allocated within the organization and it is practical to use in this system. So there is no problem for backup system and its maintenance cost is low.

The infrastructure and operation cost of IP telephony is not much expensive as PBX system or traditional system has. According to William A. Yarberry, Jr (2000), *"IP telephone can run Linux as an operating system and use off –the –shelf Hewlett –Packet or Dell boxes, the price wars begin.. 'Frank and Bill' can build a un-PBX in a garage and sell the "Frank and Bill Telephone system'' over the internet next day'* [11]. Besides, in IP telephony system data and voice can possibly send through in a single wiring infrastructure though sometimes can create problems if network failed.

2.3 Changes of IP Telephony

The rapid transition of all networks in the 21st century is digital or packet based architecture from analog or traditional network based communications. However, change has been started since 2000 where VoIP expanding dramatically and which can be integrated in all websites including email or social networking, call recording, conferencing as well as data transfer. So, VoIP established itself as a mainstream to the telecommunication industry. Skype is the best example of its type and over the time it becomes the market leader. So '*VoIP has been the harbinger of convergence between voice and data/IP networks, facilitating a growing range of unified communication and collaboration services'*. This is because of its cheapest call rate and reliability of the service. Moreover, change to IP telephony is deep-rooted because of few reasons such as; toll bypass, data voice resources, convergence everything over IP, presence over IP including voice, data and video; real time, centralized management system and remote site management. Hence, IP telephony opens broad market for vendors to provide services along with hardware, software, soft switch and router devices for business purpose. It also facilitates Skype, Vonage, ViaTalk, Yahoo Messenger, Google Talk, Jaxtr, VoIP buster to provide services. It observes that, Skype dominates the independent cheap calls VoIP markets. People are always looking for free calls if any service provider offers. Thus Skype does. That's why it is becoming the most popular in telecommunication sector in the world. On the other, Vonage policy is quite different to Skype though they have 1.6 million customers worldwide. They are not willing to give the free calls to the consumers. So the momentum of IP telephony is fast growing. Besides, Google is also planning to enter into the Mobile VoIP market.

Figure: 1 IP Telephony changes Over Time



[Source: Company websites Informa Telecom &Media, Juniper Research, IEC, Telecom Asia, News report as cited in Industry Report Volume 3, *IP telephony*]

[Figure: 1], clearly shows the changes of IP telephony and its solution for service providers.

2.4 Growth of IP Telephony over the Time

It is obvious that, new technology always takes over the old technology. So internet telephony takes over the place of circuit switched network or traditional PSTN network because of technology of choice in ages. Internet telephony offers cheapest long distance calling card and international telephone calls with much wider and diverse range of communication service for the consumer. Consequently, over the time IP telephony gained the interest of the IT and telecommunication industry, policy makers and regulatory board- as an alternative choice of IP infrastructure deployment. The International Telecommunication Union (ITU) found that, the liberalization of the market contributed a broad range of migration from traditional to IP-based network. Besides, international Public Telecommunication Operators (PTOs) have migrated all their international traffic to the IP based network and started to huge invest. Moreover, the growth of IP telephony networks is creating wider implication for the telecommunication industry along with national and international agencies and it's being viewed as a fundamental competitor among the countries in the telephony market worldwide now. It is considered as the spur of dynamic economic growth.

During the period of mid-to-late 1990's IP based network accelerated in the telecommunication industry and internet was offered over public internet like Free World Dial-up. During the year of 2000 and 2002 "VoIP was a discounted telephony over IP based networks" [12]. The best examples of its type are Net2Phone and IBasis. Over time VoIP gained popularity and become most competitive place in the telecommunication industry all over the world. Besides, Voice over broadband also enhances their service to provide free or sometimes flat rate along with the reducing calls rate to PSTN mobile users. Skype and Vonage are the best examples. Now IP telephony shows the strong market growth due to its fast growing phenomenon. Even in the enterprise voice market it grows rapidly and it includes converged IP/TDM PBX phone system along with IP phones. [13].

This paper also presents the data that, in 2009, there are 70% of the total telephony market has been represented by IP telephony. Besides, if we have a look on the vendor in the Western Europe, there is a tremendous competition observed. Jeremy Duke, a principal analyst and founder of Synergy Research Group says, *"In looking at the vendors in Western Europe, we see an extraordinary tight race for the first position. It could be argued that 3 vendors tied for the second position''* [14]. Following this [table: 1] shows the clear picture to get clear idea about it.

Table:1 IP telephony line shipment market share

Vendors	Market Share
Alcatel –Lucent (France)	17.67%
Avaya (UK)	16.94%
Aastra (UK)	16.93%
Siemens (Germany)	16.54%
Cisco (UK)	15.41%
Nortel	5.96%
Mitel (UK)	5.82%

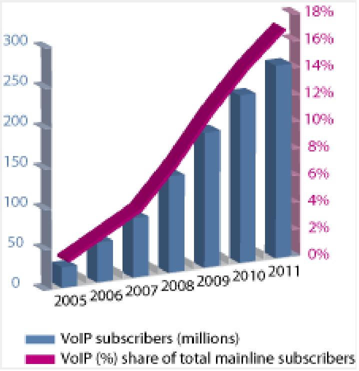
Besides, Cisco, Clarent, Nuera, Sonus, Unispere, Convergent Networks are also getting popularity for the IP Public Switched Telecommunications Networks gateway vendor. Alongside as a carrier ITXC, Genuity, Net Voice, Net2Phone, IBasis are getting place in a competitive market arena. Though there are innumerous networks operator are providing their service worldwide, Concert, PointOne, China Telecom are into the edge of the market. As a carrier Net2Phone has an excellent marketing power to control the VoIP market place and therefore they formed a strategic alliance with Yahoo, AOL and MSN.

For the growth of VoIP there are so many predictions that have been projected by the researcher. For example, "By 2008, wholesale VoIP traffic in Europe, Middle East and Asia reached 57 Billion minute [15].

On the other hand, IDATE one of the Europe's leading market analysis and consulting firm estimate of the VoIP subscribers, total and as a proportion of mainlines worldwide during the period of (2005-2011) would be 300 Million. IDC also views; IP telephony, the next generation technology and over the time it changes the telecommunication landscape where the growth is exponential. They predict; only in USA, residential subscriber of VoIP will grow from 10.3 million to 44 million within the year of 2006 to 2010.

Graph: 1 Estimates of VoIP Subscriber

Estimates of VoIP subscribers, total and as a proportion of mainlines worldwide, 2005–2011

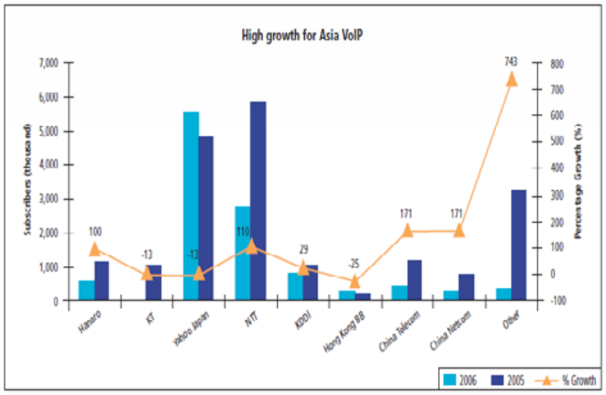


Source: IDATE as cited in ITU

Tom Evslin, the chief Executive Officer of ITXC remarks that; *"By 2010, all voice traffic will be over IP networks''* [16]. In china IP Telephony Lunched in 1999 then within short time China Unicom become the world largest VoIP network platform and the daily usage over 2.5 million. Ten million calling cards is sold and revenue comes Two Hundred Thousand US Dollar a day.

In the Asia Pacific Region Avaya Inc., is a leading global provider of communications networks and their strategy is to provide service based on innovation and IP telephony applications applied in banking sector, finance, insurance, retail, manufacturing, and travel and hospitality industries. Hence, it contributes a lot for economic growth in this region. Thus, Avaya is becoming the leader of IP telephony market in Asia Pacific Region. Recently a report published by the Asia Pacific Technology Market CY 2007 that; Avaya is in the leading position in IP telephony Market share in this area; for example, in China 30.7%, Singapore 33.8%, Hong Kong 33.7%, Philippine 57.9%, Indonesia 30.9%, Taiwan 28.8%, Thailand 29.8%, Vietnam 31.5% and Malaysia 27.7% [17]. Their main service is to focus on the customer need and demand. Avaya apparition is to make a new world of opportunities for the business and enhances customer satisfaction as well. Following this [graph: 1] shows the high growth of VoIP market in Asia and NTT is in the leading position for VoIP traffic.

Graph: 2 High Growth of VoIP



[Source: In-Stat, May 2005 as cited in Industry Report 2007, Volume 3]

The vision of NTT is to create a 'new value in communication and partner with customers to *"bridge their present and future potential''*. NTT clearing house provided around 220 countries all over the world with local access along with speedy IP telephony services. It also provides 'the single point of contact to its customers with global coverage, low cost and high quality service and makes entrance barriers to this market lower'. NTT is delivering high quality of voice, data and IP services to the service providers all over the world and building a new business model through the power of communication. Providing dedicated and quality of service NTT became the world ranked 31st in the Fortune Global, 500 list in 2010. Nevertheless KDDI, Hong Kong BB, China Telecom, Yahoo BB, KT and others also helps to expand the VoIP market in Asia.

2.5 IP Telephony Growth in Malaysia

VoIP first introduced in 1995, when internet appears into the world of communications and started to use at home basically PC-to-PC connection that are connected to the same telephony software users and connected with each other. It is free of cost that's why voice quality was very low and feedback is not satisfactory. Then in 1997, PC-to-Phone system has been incorporated with a bit improved sound quality but this method was inconvenient because of its one way of communication. After that, in 1997, VoIP or IP telephony is being introduced as a form of voice and data communication and through the government agency. Various types of service provider came into the Malaysian IP telephony market including local as well as foreign companies. VoIP market came into the highlight when Cisco System and Nortel started to produce hardware for VoIP equipment. As a result VoIP become more attractive and profitable business sector in the world. Then it implements as a main source of data communication of VoIP platform on their domestic as well as international IP networks for communication. Gradually, VoIP starts to use as a business source. In 2000, VoIP usage increased rapidly in Malaysia. Then, in 2000, there are 53 IP telephony license had been given in Malaysia but now around 400 companies are holding IP telephony license.

There are several types of services are offered in a VoIP, such as; prepaid, posts paid and call back where the target groups are foreign workers, students, professionals and immigrants.

These cards are available in different retail shops all over Malaysia. Another type is post-paid account is mainly using as a corporate account .Call shop is another type of VoIP service and the target groups are foreign workers, immigrants, teachers staff and others. This type of service also provides the long distance calling card that can be used-using local telephone number. Malaysia is strategically a good place for VoIP market, so the growth of VoIP is significantly uprising and over the time it is becoming the mainstream of economic development and incorporating the new type of communication method.

Previously, in Malaysia fixed phone telephone companies were allowed to offer VoIP service to the consumer. However, with the advancement of technology, VoIP market policy has been reviewed and extend supports to all types of service providers those are interested. There are two ways that VoIP service can be provided;

1. PC-to-PC based, that is known as an Internet Telephony and
2. Phone-to-Phone based, that is using Public Switched Telephone Network (PSTN) where multi stage access dialing are allowed. This type of service is called Voice over Internet Protocol (VoIP) or Internet Telephony (IP).

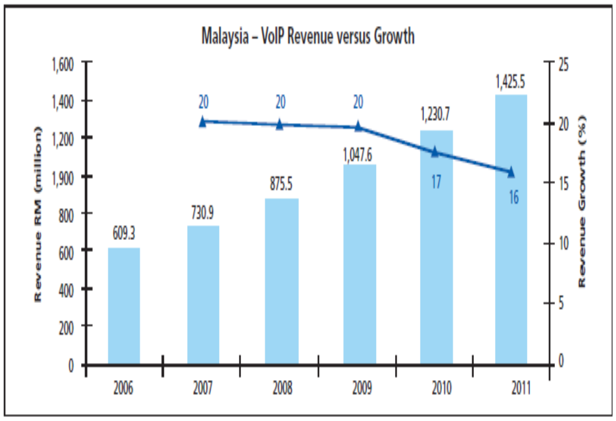
As PC-to-PC phone does not require any PSTN as a prerequisite, so regulatory commission is not imposing licensing restrictions. But, in VoIP services, regulatory authority imposed licensing order because it’s originated and terminated through the PSTN in VoIP mode. So effective from 1st April, 2002, companies those are providing VoIP service is required to an Application Service Provider (ASP) license for domestic as well as international service. Service providers are not allowed to buy minutes from overseas carrier without having NSP license or Individual license [18].

The Internet Protocol Telephony (IPT) market looks very impressive in Malaysia. IDC reported that, in 2005, local VOIP market was worth RM 645.9 million and the industry is expected to the annual growth rate is 18% for the year of 2005-2010. However, in 2005, the annual growth was reduced 19.4% whereas previous year it was 67.2%. The report says also noticed, in 2005 IP telephony service was not profitable and declining due to repatriation of foreign workers from Malaysia. Another reason is increasing competition among the incumbent service providers and decreasing call tariff rates both in domestic and international destinations. Senior Analyst of Enterprise Networking and IP Communication Researcher of IDC Malaysia, Lincoln Lee says;

“The IP telephony services market is evolving away from a discounted call services model to that of a pure IP telephony service. Service providers who do not adapt their business models to meet market demands will face further erosion of revenue and profits. Technology disrupters such as WIMAX, WIFI, unified Convergence and Mobility are acting as a catalyst of change in the IP telephony market. Growing market demands for such services and technology adoption are forcing service providers to evaluate such technology an alternative means to provide IP telephony services”.[19]

IDC made another research on the IP telephony Top Service Providers market share by revenue in financial year 2005. There were five major service provider doing their service and the contribution were as follows, Redtone 22%, Telekom Malaysia 12%, Nextel 12%, Nation Com 12%, Extive 4% and other service provider was 40%. Over the time contribution from Application Service Providers are declining. This is possibly due to strong competition among the service providers and regulatory issues. Following this [graph: 3] shows VoIP revenue versus growth from 2006 to 2011.

Graph:3 Malaysia- VoIP revenue vs growth



[Source: IDC as cited in Market and Financial Review Q3, Communications and Multimedia Malaysia]

This graph shows, Malaysian VoIP revenue vs revenue growth comparatively. Revenue growth is gradually declining compare to previous year.

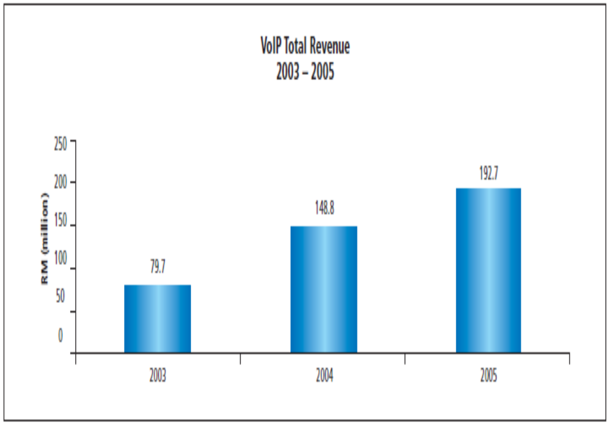
However, PCCW Global and Telekom Malaysia extended their VoIP platform for service providers and enterprise markets. PCCW is a Hong Kong based communication solutions for global businesses and service provider’s solutions robust and dedicated TDM/IP and high quality of voice termination. PCCW also serves service providers all over the world by providing domestic and international switches. Service provider’s solutions including IP and MPLS transport, satellite based video transmission, cellular backhaul, international voice termination, toll-free service and mobile messaging, high definition video conferencing, VoIP switch partitioning along with equipment services. Malaysia is getting advantages of having the regional office in Kuala Lumpur and VoIP service expanding very fast. Telekom Malaysia Berhad is the largest incumbent operators in Malaysia also providing cost effective calling card services all over the world. There are different types of card are providing to the customer such as; iTalk, italk Whoa, italk Mobile Dialer, TM calling card, iTalk international Airtime Transfer and others. iTalk have over two million subscribers worldwide. Telekom Malaysia also focused on value added service to existing and growing clients like iTalk Buddy which is highly demanding as TM claimed. ITalk Buddy users are able to send messages, make PC-to-PC calls, share files and folders, able to share blogs, upload and share photos as well. Thus an online community grows and connected with each other.

As a market leader, TM is also providing wholesale service such as; transportation, origination and termination of calls anywhere in the world. Thus, TM is working as an ideal partner for domestic network operators.

2.6 Revenue Trends

SKMM report stated that, Malaysian IP telephony services is a combination of incumbent and small players in the market and provided under Class License which has to obtain from SKMM. Licensing is required for providers under the Communication and Multimedia ACT 1998 for the provision of VoIP service. In 2003, only seven operators offered VoIP service in Malaysia and total revenue was 79.7 million whereTelekom Malaysia Berhad was the highest contributor from the incumbent player and it was 81.3%. Next year revenue grew in 86.8% and total revenue was 148.8 million. However, in 2005 revenue increases to 192.7 million, though only ten service providers offered IP telephony services. Following this [graph: 4] shows more clearly about the revenue growth in Malaysia;

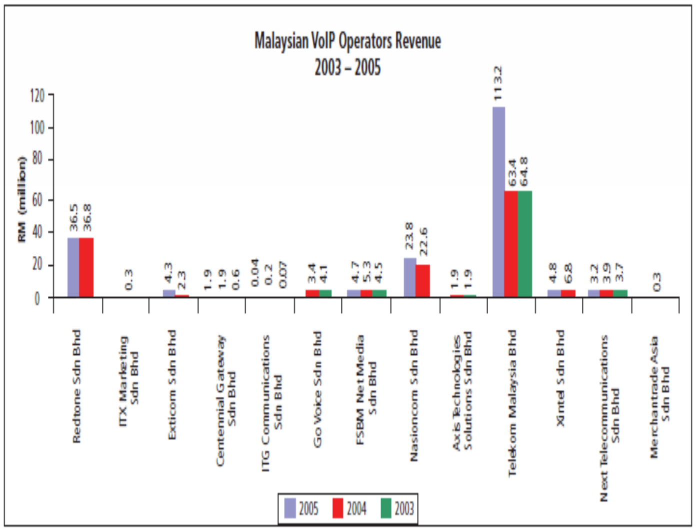
Graph: 4 VoIP Total Revenue



[Source: SKMM as cited in Industry report 2007]

Another study shows Malaysian VoIP operators Revenue from the period of (2003-2005) where TM is on the leading edge. In 2003, TM revenue was 64.8(million) followed by the year in 2004 63.4 (million) and in 2005 it was 113.2 million. But over the time, TM revenue on VoIP is declining due to the strong competition appears in the voice market. Besides, Redtone Sdn Bhd is in the second position and its revenue in 2004, 2005 were 36.8 and 36.5. [Graph: 5], shows it clearly.

Graph: 5 Malaysia VoIP Operators Revenue

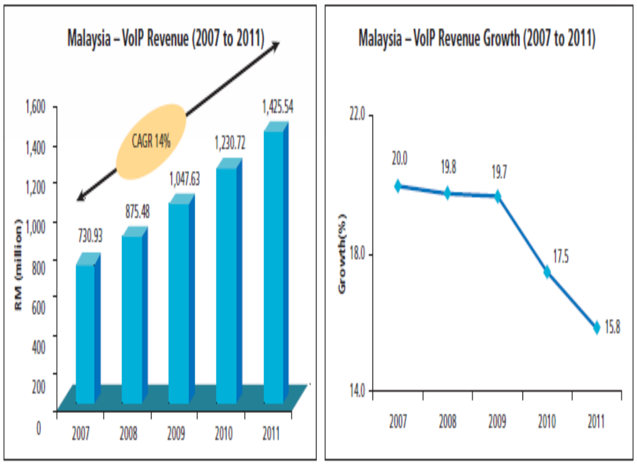


[Source: Industry SKMM, Note: 2003 -2005 revenue is based on the operators that have reported revenue as cited in Industry Report 2007 volume 3]

others service providers, such as; Nasioncom Sdn Bhd, FSBM Net Media, Next Telecommunications Sdn Bhd, Exticom Sdn Bhd, Marchantrade Asia Sdn Bhd are also contributing revenue growth. All data shows in the graph that, VoIP revenue trend is declining.TM Market share is 81.3% in 2003, 2004 it was 42.6% and in 2005 58.7% also gradually declining too. However, statistics shows that, on the following year 2006 and onwards it is promising but the revenue growth between 2004 and 2006 is only 4%. In spite of the forecasted revenue growth it is declining. There might be a reason of SKMM legislation issues or involvement of the network service operators in the local market. Previously, local incumbent operators were not interested or not aware of this type of service. Once local incumbent operators came into VoIP market the competition heats up as a trend of the voice market but foreign companies are declining edge because of not having popularity and large market coverage. DIGI is the third largest Malaysian operator has launched at a preferential rate and good voice quality on the international termination and VoIP service as a means of direct calling method. Below the graph clearly stated the forecasted VoIP revenue from 2007 to 2011 as IDC did so and in 2011 the VoIP revenue will be increasing 1,425.54 million RM as the previous year was 1230.70 million RM. Despite the forecasted growth, the revenue growth is declining year by year. In 2007 the growth rate was 20.0

and in 2008 it was 19.8. So the downturn affected to the VoIP revenue growth in 2010.

Graph: 6 Malaysia VoIP Revenue



[Source: IDC as cited in Industry report 2007]

2.7 Business Value of VoIP

If we simply think about the VoIP business value and its advantages, first thing will be appearing to us that it is the demand and the perfect time to enter into the market.

IP telephony is a fast growing and dynamic technology and it helps to access to the information age which eliminate the boundary of the communication. VoIP deployment facilitates to build the revenue growth in 21st century information age as it is new service for call transit through the VoIP gateway along with unified messaging service, virtual private network and others. It also facilitates to develop the distance learning, e-government, telemedicine as well as economic growth of the country where IP integrated network works as a catalyst for development. The most eye-catching issue of VoIP is huge cost savings for long distance calling card, new features and converged network [20]. However, the cost saving can be quantify but the productivity improvement will be very tough to enumerate. Getting a benefit from the VoIP implementation it is absolutely needed to invest in a long term period because it will "*provide returns in capital and productivity savings, and help avoid additional security risk*" [21]. Nevertheless, for productivity improvement end users and service provider needs to implement news features of VoIP, which will enable to get update and advance service like integration of Voice Mail, Email, and fax etc. that can be accessible in any places and any time. It also helps to improve the customer relationship and management system and quick virtual solution if needed.

Capital and expense saving are also important on the VoIP platform where in case of long distance calling card can vary on the distance an time period in which time end-users are calling.

The writers also mention about the productivity savings on the VoIP implementation. Thus he focuses three main issues (i) management and support savings, (ii) enhance and support mobility, (iii) and reduced site preparation time. Besides security is another challenging issue for the VoIP business implementation where security should the main concentration to prevent the server hack or damage along with service quality, user's acceptance and reliability, user and staff training on the administrative level is significant. Moreover, as the technology is always dynamic there should be the adaptation of new technology with a competent use as the VoIP market is always challenging and critical.

IP networks can be designed in order to provide the quality of service. It is consider the core of voice telephony. There are so many issues are related with the service quality including reliability and security. As Konrad L. Trope (2006) stated,

"*Managed 'IP' networks support the capability to prioritize the voice and ensure prompt and consistent communications regardless of how congested the network. In that environment the user does not distinguish a difference in quality between a managed VOIP call and a traditional POTS call''. [22]*

VoIP enhances and facilities the interconnection and eliminates the boundaries between wireless and wireline devices .even VoIP secured the geographic interdependence as Konrad stated. In other words it can be said that IP eliminates the physical boundary as well as geographic arena. So, deployment of the VoIP there is no need to set up the own dedicated telephony systems. There is only needed to access the IP WAN (Emerson). Besides, IP telephony help IT department to manage voice and data together. As it can handle from anywhere in the world or in a sense virtually it can be settled if problems occurred. IT manager can sit anywhere if WAN is available and take necessary steps for changes or solutions. However, there is innumerable business value of IP telephony or VoIP deployment there are few more advantages can get from the IP telephony services such as free inter office calls, voice mail, auto attendant, home working, flexible console, disaster recovery, database integration etc.

In IP telephony services calling card providers are fast growing. Even most of the fastest growing IP telephony carrier also provides pre-paid telephone card services. As it is fast growing there is a possibility to gain the market share rapidly and the new entrants of IP telephony services. This service is based on Packet based network helps which cost effective and scalable services, so, it brings the high revenue growth. Another important issue on IP telephony is that, ISP's service onto the calling card. Davidson & Peter (2000) says;

"Lower-cost IP infrastructures enable ISP's to pass savings on to customers in the form of lower tariffs. In international markets, where long- distance rates are high, ISP's can offer competitive services while still maintain high profits."[23]

So ISP offers cost effective services to the IP telephony platforms. In this case, end user those are subscribing into the ISP's can use Voice over Internet protocol service. On the other hand, operators may upgrade existing network based on IP networks. BT's 21st century network is the best example on this type of network.

VoIP over the internet also brings the new business opportunities which is fully internet based and users can download free voice telephony software then installed into their PC's to make free calls worldwide. Skype is the best example of this type of free software. However, PC to phone calls also possible but it chargeable [24].

2.9 Issues and Challenges

However, there are few issues and challenges of IP telephony service identified by the ITU for the developing countries are below;
– Its impact on their revenue streams, resulting from lower-priced "IP telephony" tariffs compared with their PSTN tariff schemes
– How not to place any additional requirements on PSTN networks when interconnected to IP-based networks
– How to meet the performance metrics and traffic identifications when IP-based networks interwork with PSTN
– How to generate the necessary funds to invest in IP-based networks
– How to deal with numbering and addressing issues.

2.8 Economic Aspects of IP Telephony

IP telephony implementations or adoptions have a significant prospect for the economic and social development which would lead to the country for sustainable nation. World Telecommunications Policy Forum found finds four major economic as well as social aspect of IP telephony. These are as follows;

1. By using IP – based network for electronic commerce, firms can widen their potential customer base and reduce transaction costs while national economic can benefit from new trade opportunities;
2. By using IP-based networks to retrieve information, health care professional can keep up to date with developments in specialist area and can pass on their knowledge to others;
3. By using IP –based networks as research media, schools and universities can greatly expand the range of information services available to their students and ensure that teachers remain abreast of the latest developments in their fields;
4. By using IP- based network as communication tools, governments can make their services more accessible to their citizens and can establish website to promote events provide information.

2.10 Market Drivers for VoIP

There are so many reasons to drive the IP telephony service instead of TDM service. The most important issue is cost effective and possibility of high revenue growth passes up the international regulatory fees. An integrated service offers IP telephony system. Besides, there is a possibility to outsource (VPN) which is allowed remote access service and addressed simple gatekeeper issues along with integrated voice and data services as well. In Malaysia there are various factors lead to the age of IP telephony services such as; deregulation and no barrier entry and liberalization of the market. Besides, dynamic and low cost voice service enhances to enter as an alternative voice service both consumers and the corporate market. However, as a new business model, individual and wholesale IP telephony creates new opportunity to enter into IP markets. Besides, IP telephony services offers value added services which will drive the development of IP telephony services and "*increasing numbers of service providers will be bundling together a range of IP services including VoIP with IP-VPN and other services, packaged as a total IP communications service offering'*[25]. Moreover, service providers are adopting this service because it is high revenue generating source along with value added telephony services for the SME and residential broadband service market.

3. Literature Review

This chapter reviews the literature on issues and challenges of the foreign companies those are investing in Malaysia in the fast growing service sector of IP telephony. From the trends of empirical and theoretical observation, this chapter elucidates MCMC legislation and service provider's access to the market which brings difficulties to implement and run the IP telephony service in Malaysia. It examines also the role of TM and restriction issues along with pricing challenges. In the competitive and fast growing telecommunication market, Internet Protocol or VoIP is being enormously adapting for transferring the data and voice communication. It will also examine pricing issues of PSTN and VoIP in Malaysia and pricing differences with other countries. This chapter illustrates a depth analysis of licensing restrictions and Malaysian regulatory issues which affect IP telephony market and its growth.

Having low cost facilities and favourable investment opportunity, IP telephony business structure helps to generate more revenue sources. It creates opportunities to consumers, enterprises, and SME's an alternative business field.

The telecom industry critical regulatory dilemmas are somehow badly affected in IP telephony market. So this paper analyzed the VoIP network access with legislation issue. In telecom industry critical regulations issues hindered the growth of world IP telephony market.

In this area regulation is the main driving force as it tightly regulated[26]. Problems are arisen due to regulatory attempt especially in the developing countries, where IP telephony leverages the prospective of the telecommunication sector for data and voice transmission. Besides, strong competition appears between incumbent and entrant where incumbent is balancing with PSTN and VOIP telephony and entrant is struggling for market entrance where regulatory issues intervenes or restricts upon their services [27]. Legislation issue supposes to appear to put off the anticompetitive attitude towards the service but not for prohibition. If it happens in other ways, innovation and development process will be affected. Refers to Ebril and Slutsky, 1990; and Lewis and Sapington 1990, Paul de Bejl & Martin Peitz remarks;

"If in a particular, the regulator can set different rates for a bottleneck owner and a non-integrated competitor; the regulator may want to subsidize the competitor at the margin to increase competitive pressure".[28]

Due to the access price of VoIP, ASP license holder's especially foreign entrants are facing severe problems to launch their services and confront challenges with local network service providers in Malaysia.

Global access price for the entrant should similar with incumbent operators. If not then, only existing market players will get benefit and no market entrants will survive in this market especially in Malaysia, because incumbent operators has a large local and international subscribers. On the other hand, large volume of usage is another reason for growing market as price is determined by the volume usage.

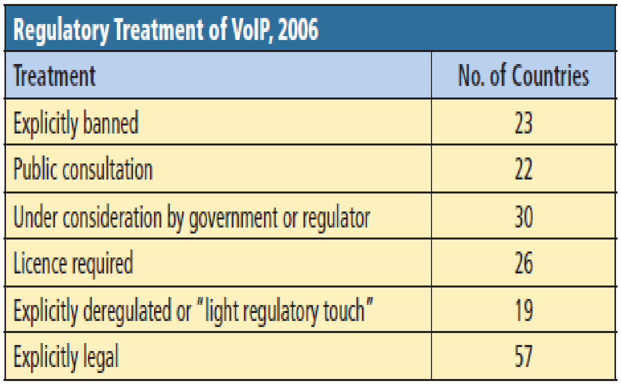
Nevertheless, regulatory issues can be imposed in the retails IP telephony market to foster competition. But incumbent operator or the existing network service providers gaining market advantages in every aspect and getting privileges when they provide services.

3.1 Regulatory Issue and Market Information Worldwide

In IP telephony services, regulatory issues varied country to country. The purposes of regulatory issues are to intervene or to control or to regulate the services. However, regulatory issues also use to ban or disallow the service in different countries in the incumbent market. ITU Telecom Regulatory questionnaire shows, fifty seven countries are allowed to provide VoIP service, and 26 countries are required to have license to provide the service and 23 countries banned VoIP service explicitly [29]. In a number of European countries, VoIP is seen as a *"light regulatory approach"* where they viewed, VoIP is not a main part of the telecommunication network and data service over internet are mostly unregulated [30]. The *"light Touch"* approach of EU regulation is using to expand the broadband access and persuade the competitive attitude between traditional carriers with IP based carriers.

So, regulatory approaches are helping to promote rapid innovation and determined the customer's interest. So, distinctive telephony regulations apply in different countries based on the type of services is offered. ITU report shows that, most of the national policy of IP telephony focuses phone-to-phone service. Table: 1 shows the regulatory treatment of VoIP worldwide.

Table: 2 Regulatory Treatment of VoIP



Regulatory Treatment of VoIP, 2006

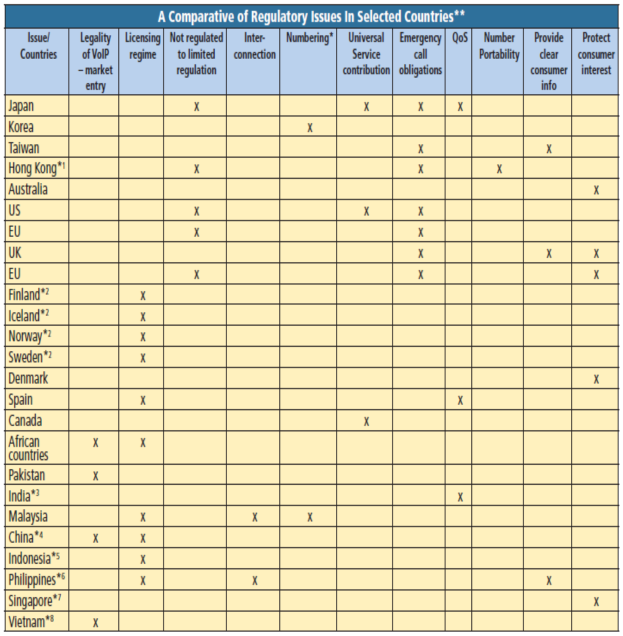
Treatment	No. of Countries
Explicitly banned	23
Public consultation	22
Under consideration by government or regulator	30
Licence required	26
Explicitly deregulated or "light regulatory touch"	19
Explicitly legal	57

[Source: ITU telecom regulatory questionnaire 2006]

On the other hand, 'PC-to-Phone services tend to be prohibited in those countries that prohibit IP Telephony generally, while they tend to be permitted without condition in countries that permit some or all forms of IP Telephony' [31]. In Australia there are no specific policies that can regulate IP telephony or VoIP though Australian government has *"long considerate loosening their licensing regime to encourage broadband network rollouts and increase consumer take up of VoIP"* [32]. In 2008, $412 million achieved at the VoIP services in Australia and $187 million revenue accounted in the residential VoIP service sector. Compare to business VoIP, residential VoIP is not fast growing where business VoIP services grew to $225 million where 270 companies are providing their service in Australian local market. One of the famous Research group Market Clarity estimated that, from July, 2007 to June, 2011 VoIP subscribers will be increased 1.4 million to 4.8 million. In local SME, VoIP have good opportunities to extend their market for business solutions. A recent survey conducted by Sensis that, only 13% SME are using VoIP service but within the twelve months' time, 70% of total market will be VoIP marketplace. So it is identical that, the expansion of VoIP in local as well as foreign market will be covered a huge area because of its easy access system.

Like Australia; Japan, Korea, Taiwan, Hong Kong, US, UK, EU and Singapore didn't impose strong regulation on VoIP services. As a result VoIP is fast in these countries and become the leader of the world market. In Asia Pacific region Japan, Korea and China expected higher VoIP growth. In 2007, there were 21 million subscribers but at the end of 2011, subscriber will be 42million [33]. Countries like Finland, Iceland, Norway, Sweden, Denmark, Malaysia, Spain, China, Indonesia and Philippine are required to have license to enter into the VoIP service. In Japan, there are no specific laws for VoIP. IP telephony has become the preferable voice service and estimated that, 80% of total VoIP subscriber of the world in Japan. The VoIP market in Japan is much competitive as the market is privatized. Best example is Nippon Telegraph and Telephone (NTT) whose main slogan is "Creative life for everyone''. It gives access to its high speed internet network along with related feasible supports to its client. Due to the lower cost, IP telephony market changes considerably and market size has increased enormously. The best examples are KDDI and NTT. In 2007, KDDI replaces all the networks to IP telephony structure and NTT replace 30 million of metal subscriber's line to FTTH by the year of 2010 [34]. VoIP market stands astride in between traditionally regulated and relatively unregulated voice and data service market in EU. Presently, EU telecom regulators are grappling the challenges of regulated VoIP market. Besides, European Commission divided VoIP service into four categories. These are; PC-to-PC where everyone uses software based system; VoIP within private corporate network; public operators use of VoIP over the PSTN traffic on their core

Table: 3 A comparative Regulatory Issues



A Comparative of Regulatory Issues In Selected Countries**

Issue/ Countries	Legality of VoIP – market entry	Licensing regime	Not regulated to limited regulation	Inter-connection	Numbering*	Universal Service contribution	Emergency call obligations	QoS	Number Portability	Provide clear consumer info	Protect consumer interest
Japan			x			x	x	x			
Korea					x						
Taiwan							x			x	
Hong Kong*1			x				x		x		
Australia											x
US			x			x	x				
EU			x				x				
UK							x			x	x
EU			x				x				x
Finland*2		x									
Iceland*2		x									
Norway*2		x									
Sweden*2		x									
Denmark											x
Spain		x						x			
Canada						x					
African countries	x	x									
Pakistan	x										
India*3								x			
Malaysia		x		x	x						
China*4	x	x									
Indonesia*5		x									
Philippines*6		x		x						x	
Singapore*7											x
Vietnam*8	x										

- *Include specific numbering using and geographic numbering
- ** Indicates reference in the country's context
- *1 Plus back up power supply for "lifeline" devices
- *2 PSTN regulations
- *3 Distinction between toll quality and below toll quality in Nov 02
- *4 Skype is deemed illegal
- *5 Consideration to re- formulate regulatory issues like numbering, access code, routing and interconnection when VoIP growth goes from service based to facilities based operator
- *6 VoIP is treated as a value-added service; registration required but not authorization; interconnection intervention If need be by regulator
- *7 Technology neutral approach; no obligations except to inform users of service information and clear information about service capabilities (emergency service access and quality)
- *8 Subject to price regulation

[Source: "The Status of Voice over Internet Protocol Worldwide", the of Voice Workshop, 15-16 January, 2007 as cited in Industry report 2007] network; fourth is publicly available VoIP services which covers by the New Regulatory Framework for Electronic Communication (NRF), though they are not still decided how to apply rules upon it.

In Finland, Telecommunication Market Act does not cover voice transmission in a data transmission network as it is not considerable part of the networks functions. Thus, they don't have regulation on transmitting voice and data over IP based network and this service is free for all internet operators. In near future, they don't have any plan to impose regulatory issues on it. However, in Germany, 'VoIP, with its different technical possibilities (PC-to-PC, PC-to-Phone, and Phone-to-Phone) is seen both as a telecommunications service from the technical point of view and, so-called tale-service from a content point of view' [35]. Nevertheless, VoIP is still undecided to identify as voice telephony and would be controlled by the Telecommunication Act. It is allied with competitive carrier in the world market and French Telecom appears as a largest consumer VoIP provider in Europe even though British Telecom and Telecom Italia are considered the top ten VoIP operator in Europe. VoIP market in Europe is highly diverse and remains fragmented due to its different regulations along with wide range of business model adopted in different types of service providers. VoIP adoption differs from country to country. For an example, there are 34% household subscribed VoIP, though they are new entrants for providing the IP telephony or VoIP service [36].

In Canada, VoIP is allowed and declared by the Canadian Radio Television and Telecommunications Commissions (CRTC) in May 2005 that, the regulatory issues will be impose "only when it is provided and used as a local telephony service'' [37].

In Hong Kong, office of the Telecommunication Authority regulates the Telecommunication Industry but there is no regulation of VoIP technology. So, it assumes that, "Hong Kong Government is technology neutral'' and it does not favour any sorts of telecommunication technology. In 2004, the decision of OFTA that, not to impose levy in local access on VoIP calls. Providing VoIP Service in Korea, there is no license required from the regulatory authority. Thus VoIP is booming in Korea overtime. The Korean Times reported that, VoIP subscribers are increasing rapidly, and at end of the 2009 subscribers increased 5 million, but in 2008 it was 2.5 million. KT Corp is the dominant traditional fixed line operator in Korea. At the beginning they were unwilling to migrate from PSTN to VoIP. As a result, it was affected severely in their business where revenue declined. Finally they felt, here is way to improve the revenue without VoIP, and then finally embrace VoIP to survive and compete in the market.

Though, VoIP regulation and the policy are distinct from one country to another, unregulated country are in the competitive edge comparing to the regulated country. Besides, an efficient regulatory regime can assist to develop the IP telephony service for all communications based service including voice and data transmission.

3.2 Malaysia Legal Framework

Though IP telephony or VoIP service proliferate rapidly in the world, regulatory debate came into the spotlight nationally or globally. There are a large number of issues raise in the VoIP service that are significant such as; allowed to provide IP telephony service or not, what type of regulation will be imposed, market entry barrier, customer protection, privacy and technical safeguard, technical attribute of VoIP and others. At present, there are large numbers of countries unregulated VoIP services and few others countries imposed regulatory issues similarly to PSTN regulation on VoIP.

Based on the regulatory regime of relevant to internet service, Communications and Multimedia Act 1998 and Communications and Multimedia commission Act (1998) are the main source for legislation. So, Malaysian Communications and Multimedia Commission (MCMC) is the regulator for the converging Communications and Multimedia Industry and Communication. Besides, Multimedia (licensing regulation 2000) and Communications and Multimedia (licensing) amendments regulations 2001 are issued as a subsidiary legislation. The primary role of the Communications and Multimedia Act is to implement and promote government's national policy objectives for the communications and multimedia sector. The commission is also issued the new regulatory framework which has been categorized as social, economic and technical regulation and consumer protection. These all are used for converging industries of telecommunications, broadcasting and online activities. Economic regulation takes an account of the promotion, competition and the prohibition of anti-competitive conduct, enforcement of access codes and standards along with the *"licensing, enforcement of license conditions for network and application providers and ensuring compliance to rules and performance/service quality''* [38].

The Ministry of Energy Communications and Multimedia had issued a policy that, PC-to-PC phone is not subject to regulate for licensing. However, VoIP service provider is required to have an individual license under the Communication and Multimedia Act1998 for the provision of VoIP service. Besides, the existing telecommunications operators are allowed to provide VoIP service as it issued under repealed Telecommunication Act 1950. According to the pursuant to Section 44,126 and 127 of the Communication and Multimedia Act 1998, there are several types of services which are, Application Service under ASP license and Application Service under Class License are being implemented and subjected from the first April, 2005. In respect of the provision of Application Service Providers, following these services is offered to the incumbent and entrant operator. These are;

i. PSTN telephony,

ii. Public cellular services,

iii. IP telephony,

iv. Public pay phone services,

v. Public switched data service,

vi. Audio text hosting services provided on an opt in basis,

vii. Directory services,

viii. Messaging services, or

ix. Such other applications service are not exempted under the Act or not listed in this sub regulation [39]

3.3 Reason for Licensing and Its Framework in Malaysia

Communication and Multimedia Act 1998 is basically undertaken the activities which are market oriented and creates opportunities for expansion the market area. Under the Communication and Multimedia Act, four categories of license are activated for regulating the market. These are NFP, NSP, ASP and CSP. Within four categories of license, two types of license are given for business, one is Class license and another is Individual license where individual license is required with very strict criteria and controlled by regulatory framework. However, four types of service providers offering distinct type of services. For example, satellite earth stations, broadband fiber optic cables, telecommunication lines and exchanges, radio communications transmission equipment, mobile communication bases station and broadcasting transmission towers and equipment are being offered by the Network facilities providers. Network Service Providers offers services like, basic connectivity and bandwidth and it helps to connect different networks for enhance the service. Furthermore, in ASP several types of function is allowed such as data and voice transmission, content based service and others for end-users.

The purpose of the licensing regulation is to promote fair competition as well as enhance the market development process where regulatory authority will monitor all activities like transparency of the services. If anything needs to solve or any new resolutions appears or any other disputable issue comes, licensing authority will solve those. As it is said, service provider and licensing authority would be transparent, so, it will help to form effective regulation and monitor all the matters relating to service and performance indicator submit to the Ministry at the end of each financial year.

So, the issue of market stability and sustainability come across with the licensing issues. If rules create only for imposing without observing the market, and giving priorities of service providers stability it will be vain at the end. Besides, high degree of regulation control possibly affected badly for the market growth and economic development.

3.5. Licensing Restrictions and Market Fostering Framework:

Licensing is one of the key factors for the telecommunications authorities in IP telephony service providers worldwide. So, licensing can be either presumes prohibiting or permitting the service that offered for the new market entrants [40]. This section tried to bring out the worldwide licensing regulation and restriction especially in China, Australia, UK, Malaysia and Japan and the way they enter into the competitive edge in VoIP market. VoIP growth is descending overtime in Malaysia but Japan, Australia, China and France are ascending and leading the market. In VoIP services, rules and regulation could be dynamic and market oriented. In order to protect ASP holders from the monopolized market, MCMC can take initiatives market-driven approach as Japan follows currently where no electronic surveillance and VoIP service provider is not subject to buy the volume minute from any local incumbent operators.

Experience of Japan shows with the government initiatives that, how much market progress achieved where policy is observed as an interactive ‘guiding process’. Without government initiatives new technology can be proliferated significantly as a substance of innovation process. Unlike other countries, Japan does not have legal challenges of VoIP service. Before 1 April 2004, telecommunication authorities categorizes two types of ‘telecommunication business law’ such as; type 1 and type 2.Type 1, does have own facilities to provide VoIP and type 2, does not have their own facilities instead of leasing their lines to provide services [41]. Later on in 2004, few amendments come out into Japanese telecommunication law which is stated below;

"Any person who intends to operate telecommunications business by installing telecommunications circuit facilities on a scale exceeding the standard specified in the applicable Ministry of Internal Affairs and Communications (MIC) ordinance shall obtain registration from the Minister for Internal Affairs and Communications'' [42].

Any person can apply to get a license and can offer IP telephony service as well. However, in Australia there are no licensed required for service providers but for the carrier service must have individual license. The Telecommunication Act 1997 stated that,

"*The regulatory framework under the Act [The Telecommunication Act, 1997] differentiates between service providers and carriers in terms of their legislative rights and obligation'' service providers are not subject to any licensing requirements but are required to comply with a range of obligations including standard service provider rules set out in Schedule 2 of the Act''*. [43]

Like Japanese legislation, in Australia any person or corporation or partnership can apply for license to provide VoIP service. According to the MCMC rules, any public body or joint venture with the proportionate of securing local share holder along with confirmation of buying volume of minutes from any local network service providers are acceptable to buy minutes volumes. It is well known that, due to the fewer infrastructures cost, VoIP pricing is always less than PSTN pricing. But if it is being imposed to buy with high price from local incumbent operators definitely Application Service Provider will no longer sustain in their VoIP service market. However, in Singapore, the Telecommunication Act where in Section 5 (1) and 5(2) issues license to "…any class of persons…." an provide the service and can get a class license [44]. In Malaysia,

"…both the Malaysian Communications and Multimedia Act 1998(the MCMA) and the Communications and Multimedia (licensing) Regulations (the 'MCM licensing regulations'') contain important information about the regulatory framework for the issuance of the registration for class licenses. For example, regulation 17 of the MCM licensing regulations provides that the minister may decide that a service or any activity will be subject to a class license. However, the minister's authority to make such a determination is established in section 44 of the MCMA''. [45]

In addition, there is no licensing regime for VoIP service providers in UK, but in July 25, 2003 has implemented a new EU framework for the directive of electronic Communications network and Service Providers called the Communication Act, 2003.

"The Framework sets out a harmonized and technology neutral regime for the regulation of communications companies across the EU, which will provide industry with greater certainty and a transparent more uniform approach across the members states .the regime is based on five EU Directives that cover interconnection and access, data protection, universal service, authorization of electronic communications networks and services and a common regulatory framework. the requirements of four of the Directives have been taken forward in the communication Act 2003, and following the enactment of the communications Act and the change in regulatory regime certain parts of the Telecommunications Act 1984 have been repealed.'' [46].

So, the regulatory process is transparent and neutral. Regulatory towards four of the directives in EU and the regulatory authority Ofcom sets seven principles which are most effective for policy implementation and market growth. If market cannot achieve alone, Ofcom will intervene towards the public policy goal and will operate a bias against market intervention. But, if any case, intervention needs to impose, regulatory authority will come out with a firm decision where applicable.The most efficient principle of Office of Communications (Ofcom) is about the intervention which is "*evidence based, proportionate, consistent, accountable and transparent"* [47]. This type of regulation undoubtedly fosters the market growth because of its practical implementation and effective use of the principles. In a sense, there should be the monitor groups of market observation that, if market fall down due to regulatory reason; so it ought to change where applicable. Otherwise market cannot be achieved alone and foreign joint venture will be facing problems of continuing IP telephony service in any place in the world. In Malaysian context there are no regulatory principles have been set to protect the market entrant and existing service providers. As Raslan says,

'There is no specific authority to in Malaysia to regulate competition or anti competition conduct general in Malaysia. But the MCMC empowered to regulate competition and market conduct in the telecom and broadcasting sector' [48].

As a result market becomes monopolized and Application Service Providers are not getting the benefit from IP telephony service. As a result, contribution of foreign investors in this sector is decreasing over the time and switching to other business.

3.5 Private Sector Participation

According to the ITU report, there are few category have been identified where the highest participation is allowed. These are facilities based operator, spectrum based operator, local service operator, long distance service operators, international service operators, value added service providers, internet service providers and others categories. Among the Asian countries Japan, Jordan, Pakistan, Cambodia, Bahrain and Singapore allowed 100% foreign participation in all types of sectors. Besides, Thailand and china allowed 49% foreign participation but only for internet service providers sections China allowed 50% of foreign participation. In Korean Republic 49% foreign participation is allowed but in case of value added and other categories they allowed 100% participation. Nevertheless, considering on the domestic regulations, 'Malaysia has reclassified its communications service sector on the basis on provision of Network Facility Service (NFP), Provision of Network Service (NSP), and Provision of Application Service (ASP). For NFP (Individual) and NSP (Individual), foreign shareholding is up to 30%. For ASP's foreign shareholding of up to 49% is allowed' [49].

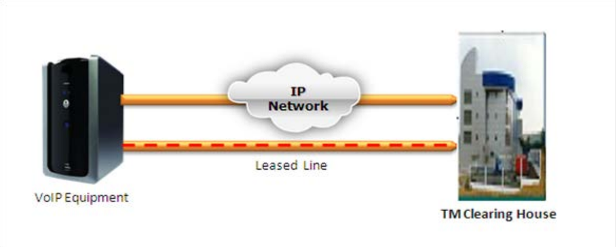
3.6 Role of Telekom Malaysia

TeleKom Malaysia Berhad is the incumbent telecom operator in Malaysia where Malaysian government is carrying extensive share 43.25% [50] through various agency [51] and facilitating the key figure of the VoIP termination in Malaysia which is being proposed as a prioritized termination, though there are few network service providers exists in Malaysian Telecommunication market. Applications Service Providers (ASP), Network Service Providers (NSP) and Network Facility Provider (NFP) are being purposely continuing the IP termination though TM Clearing House (TMCH). Meanwhile, Application Service Providers are directly controlled and monitored by the regulatory authority in Malaysia. But for if service providers have Class License, Application Service Provider may allow terminating their calls from outside Malaysia but Providers are bound to buy volume minutes from at least one local service providers though the minute volume price is higher than the global carrier. So as defining the VoIP IP termination by *TM,*

…… "Is a service that enables the Customer to terminate their fixed and mobile voice traffic to worldwide destinations via TM Clearing House (TMCH). At present, TM has two platforms of TMCH, that is VoizBridge located in Putrajaya, and Nextone, which is physically located in Berjaya'' [52].

Bellow the diagram [6] gives the clear picture of VoIP IP termination through Tekekom Malaysia Berhad.

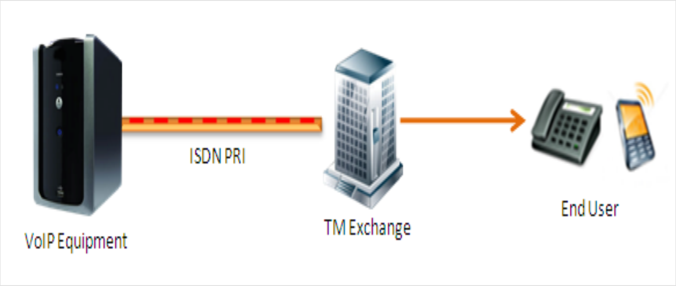
Fig: 2 VOIP IP Terminations



[Source: TM website]

Besides, for VoIP premium termination there is needed to pass through ISDN PRI and the diagram shows clearly.

Fig: 3 VOIP Premium Terminations



[Source: TM website]

From these two diagrams it is clear that, there are two types of route are to use for call termination, one is Premium route and another is VoIP route and there is a choices of being connectivity with TMCH such as, AIMS POP, leased line and public internet**.** There are several types of benefits have been identified for VoIP IP termination which are fully dedicated and extensible use all over the world by TM. These are as follows;

i) Extensive network of providers and partners located all over the world.
ii) Ability to adapt cost-saving measure by having two routing options (premium/VoIP route)
iii) One –stop solution for customer to gain access to worldwide call termination.
iv) Extensive network of providers and partners located all over the world

v) Ability to adopt cost-saving measure by having two routing options (Premium/VoIP route)
vi) One-stop solution for Customers to gain access to worldwide call termination [53]

On the other, previously mentions above that, there are two types of termination are currently available for the TM. So VoIP Premium Termination allows the Customer,

….. *"To terminate their fixed and mobile voice traffic to destinations within Malaysia via TM's ISDN PRI. The service allows Customer to utilize TM's PSTN/NGN infrastructure to make premium domestic telephone calls, hence ensures a high quality of call and is equivalent to normal PSTN telephone calls"*[54].

According to VoIP Termination rules, IP telephony International access is not allowed. Even incoming to ASP server also prohibited. Besides, VoIP IP termination towards international is allowed or in another way having VoIP IP termination permission incoming to TMCH as well as outgoing to TM fixed line, TM fixed CDMA, Mobile and international access can be possible through TM.

These issues are being appeared as a main focus point because of ASP service Providers are needed to require individual license if they want to start their service internationally. Otherwise they are not allowed to enter the VoIP market to provide IPT service. IP termination, redundancy and quality of the service are questionable. Above and beyond, the price rate of IP termination is high and service providers are not willing to buy from TM because of high price and lack of rapid technical support when necessary.

3.7 Pricing Issues on PSTN and VOIP Service Provider

Pricing issue on PSTN and VoIP service providers plays crucial role as it differs one to another and in IP telephony network it is very difficult to categorize the cost on local or long distance call because of its connectionless and in some cases "*Closed User Group*".

Though the price of PSTN termination and origination is less than the IP termination but PSTN costing and pricing model is more efficient and widely used. Unlike PSTN pricing and costing where price is determined by the miles and minutes, IP telephony pricing mostly determined by the Minutes of Use (MOU), or volume usage which can turn into the leading service provider in anywhere in the world.

In general, IP telephony is fully competitive and market driven forces. There is an important role for the regulatory interventions on price to give to motion of competition and protect the market both in local, domestic, long distance and international calling cards or call back cards. Beside, tariff rebalancing also may plays the vital role for market growth as Hong Kong adopted in their country. The Essential Report on IP Telephony stated that;

……. "A market driven approach whereby individual public telecommunication operators would, according to their business incentives and market competition, develop and implement the required IP telephony services and networks to cater for the market demands''[55].

For the tariff rebalancing approach India Telecom Regulatory Authority practices the "*transparent Tariff Fixation"* practice to fix the tariff for different service providers. So;

"The basis of fixing such charges was the underlying cost of the network elements involved in setting up for a local call, a national long distance call and an international call, in addition to cross- subsidization. Since the network elements are fixed and identifiable as Local Loop (LL), Local Exchange (LE), Transit Exchange (TE), Transmission System (TS), etc., such an exercise has been relatively simpler" [56].

However, Cross Subsidized of pricing for long distance and international call for service provider there might be bad effect on the market and competition would be decreased. Even without the prior notice Service providers both in entrant and existing may disappear on the VoIP market, because one group pay relatively low price and another group pays high price of minutes volumes for termination. Due to strong competition of this service it will be tough to get even marginal profit if service providers buy with high price. The efficient and liberalized market entry and no cross subsidy's availability can foster the market growth in Malaysia.

Further, this price differentiation may create dilemmas to the Service Providers as well because cost may hinder market share and market expansion.

3.9. National Policy objective:

The significant changed has been go through in Malaysia when the Communication and Multimedia Act (1998) (CMA) appeared in telecommunication, broadcasting industry and computer network. The Act established the regulatory framework to give supports to reach the target of the "*global centre and the hub for communications*" of the National Policy Objectives [57]. CMA plans to set a common regulatory provision to foster the market growth. Besides, building a development nation status within the year of 2020, Malaysia initiated and planned to implement equitable digital opportunity as internet is commenced in all areas of communication. At the same time, expanding the growth of ICT and Multimedia industry, Malaysian Communication and Multimedia Industry set up the Ten National Policy of objectives. These are;

1. 'To established Malaysia as a major *"global centre and hub for communications"* and multimedia information and content services'
2. 'To promote a civil society where information based services will provide the basis of continuing enhancement to quality of work and life';
3. 'To grow and nurture local information resources and cultural representation that facilitates the national identity and global diversity'
4. 'To regulate for the long term benefit of the end user'
5. 'To promote a high level of consumer confidence in service delivery from the industry'
6. 'To ensure an equitable provision of affordable services over ubiquitous national infrastructure'
7. 'To create a robust applications environment for end users'
8. 'To facilitate the efficient allocation of resources such as skilled labour, capital, knowledge and national assets'
9. 'To promote the development of capabilities and skills within Malaysia's convergence industries'
10. 'To ensure information security and network reliability and integrity'. [58]

These national objectives strengthening the pro competition, then it allows for the direct competition and it is technologically work as neutral.

"The CMA also aspires to flexibility and contains few definitions and few proscriptions. IT therefore enables ongoing reform without changes to the legislation as the implications of a converged environment emerge and evolve" [59].

Though CMA addresses the competition in the marker, it did not introduce liberalization. On the other, in 2002 report, the foreign investors, they are allowed to be the ownership as a percentage of 61% while 49% will be reverts after five years. In the context of market entry and growth, Malaysia opens IP Telephony Market earlier than any other Asian countries and became the ICT hub in developing countries.

4. Research Methodology

The research goal is to discover new knowledge of IP telephony and different regulatory issues worldwide along with Malaysian IP telephony licensing regime and its solution. IP telephony fosters the economic growth and explores new opportunities and protecting the existing market. This research will explore 'solving problems' method by exploring new dimension of licensing regime. In this regard both primary and secondary research has been conducted.

The location had been chosen Malaysia because, strategically its market place is robust, competitive and dynamic where multicultural and multi types of working forces are living in Malaysia. During the research, the researcher did not get copiousness of materials though the researcher had collected materials from MCMC, IDC website, ITU website, International Islamic University Malaysia library, company published information, Application Service Provider's information, online journal, Telekom Malaysia Berhad website, internet access as well as relevant IP telephony service provider's empirical documentation.

This study has conducted also the participation of different IP telephony providers including TM and MCMC licensing and universal service provision department. In this research, data has collected both individual and organizational level. Besides, this research also conducted interview of the TM Wholesale carrier service manager and SKMM licensing department. However, the research tool also gives strength to the researcher focusing the primary research which is basically based on interview and the secondary research focusing on the textual analysis, journal and articles, documentation and report analysis along with practical observation of working place and the survey of IP telephony Company. The reason of these data sources to observe market growth and real condition of these companies. So, the research is based on "what is observed" and its difficulties as well as market opportunities. Practical observation and document analysis reveals the reality of the research.

4.2 Primary Research

The researcher visited in different sectors of Telekom Malaysia Berhad especially in wholesale Carrier service department, The manager of ASP/ISP in Telekom Malaysia Berhad, TM wholesale billing management, licensing department of the Malaysian Communications and Multimedia Commissions, Universal Service Provision Divisions of MCMC, Aims Data Centre, Global Transit Communication Sdn Bhd, TIMEdot com Berhad and few others Application Service Providers corporate offices. Due to the inter-link with each of the company researcher had to visit them. In order to provide VoIP service agreement is pre-requisite with TM wholesale carrier, for international and domestic termination via THCH normal and premium route. TM provides ISDN PRI for international and local access. However, TM testing procedure is too long and sometimes access route testing is not satisfactory because of low sound quality. Another issue of testing is only performing during working hours and availability of the TM technical Team. Installing toll-free is a crucial issue for the service providers that, if the redundancy of calls is not good and if interrupted then it will create a severe problem for the service providers.

4.3 Secondary Research

When the researcher conducted this research, extensive assistant has received from the MCMC library where innumerous data had collected. So, whole scenario of IP telephony market is reflected on this research. Even, extensive knowledge of market strategy and the growth of IP telephony market in Malaysia have reviewed.

As a secondary data, different types of book related to pricing issue, licensing regime, growth of VoIP along with MCMC journals and articles, internet search, confidential documentation analysis and practical observation help to proceed for further development of this writing.

There are several books of IP telephony about VoIP deployment and the regulation worldwide extends the area of this topic. The book *"privacy in Electronic Communications: the regulation of VoIP I the EU and the United State"* published by 2009 and few more books assisted to know about depth knowledge of the regulatory issues worldwide and differences from one country to another as well. Writing this paper, Industry report, role of MCMC statistics report and how regulations evolve over the time has given wide area of analysis. In this regard "IP Telephony Industry Report, 2007" along with quarterly Published report on "Selected Facts & Figures'', Q1 (2010), Q1 (2007), Q1 (2008) Q2, (2008) which are focuses on Malaysian ICT indicator, broadband usage, penetration rate, national policy objective and the communication and Multimedia helps a lot. Unfortunately, in the last quarterly, 'Facts and Figures' did not focus growth of IP telephony and number of license increase; rather it focuses the basic indicator of Malaysian ICT and licensing information. Nevertheless, "Market & Financial Review" Q3, 2007 provides statistics report and comparative analysis with other Asian Countries. Another significant report, on telecommunications which has been initiated and published by the International Telecommunications union "World Telecommunication Policy Forum (WTPF 2001) Geneva 7-9 March 2001 help researchers to broaden the knowledge of telecommunication law worldwide and VoIP status all over the world.

Besides, IDC report provides the snapshot of Malaysian as well as other countries VoIP market growth and status. However, Business Overview of VoIP in Malaysia, Vocial Scape Report (November 2004) also helps to conduct this research.

5.1 Data Collection Procedure

Time constraint has affected a lot to get more access and collecting data. Researcher's practical experience and different interviews along with official meeting of these issues helped a lot to expand knowledge in this field. Moreover, co-operative attitudes of interviewees from different companies are substantial source for the research. However, the researcher's past working experiences of Japan based Application Service Provider's joint venture helps to get details information and few questions need not ask for conducting the research.

The researcher takes note while interview session held on. The starting points of data analysis are from primary sources like TM, MCMC and other involving companies. However, the researcher conducted interview at licensing department Malaysia, about licensing procedures for the Application Service Providers, return of net revenue and other issues. Here researcher managed to get fluent and co-operative response.

On the other hand, researcher's direct involvement or access through the root and confidential level of data further helps to proceed. In the process of ongoing topic, researcher took face to face interview of different Application Service Providers Managing Director, CEO, company owner and Marketing Manager. Besides, the researcher also got chance to ask random questions of related to the objective, market analysis and research questions.

There few others interview and practical observation was done in the issue of wholesale carrier and toll-free access at TIMEdot.com, sells department in Global Transit Communication and Aims Data centre for dedicated Internet Service.

As toll-free is important for Application Service Providers, so providing service, easy access and redundant channel are important. Toll-free number is needed for calling card solutions which could be reachable from any telephone booth, public phone, mobile, land phone, university booth through U1 channel. The interesting issue of the service provider's expectation of toll-free is to get access from mobile but TM is not offered for the Application Service Providers.

According to the practical experiences of IP Telephony Service Providers that, getting free access from mobile toll-free will gain market advantages. They remarked that, In Japan toll-free is free from any public booth, telephone, mobile or any other types of calling system. As a result, they are doing well in their business and continuation of business is very smooth and profitable. They also expressed their thought that, there is a possibility to switch their business off in Malaysia and start in Japan.

5. Data Analysis and Research Findings

Data is analyzed visually related to the findings and solved research questions of IP telephony in Malaysia. Researcher approaches to analyze data in a chronological sequence where practical observation, answer to open ended questions, face to face interview session, licensing authoritative body, IP telephony providing companies specially, the foreign joint ventures with a broad coverage of long distance calling card, call back, and call shop solutions are focused. Besides, other sources such as; text books, statistics reports, annual report of IP telephony, facts & figures, graph, articles, documentation, journal, ITU website, TM website and various IP telephony service providers website helped to analyses data. During the research it is very obvious that, in IP telephony services-Telekom Malaysia Berhed and Malaysian Communications and Multimedia Commissions are the main authoritative body.

The most problematic situation that faced entrants in IP telephony market is licensing issues. If the interest foreign venture wants to invest in IP telephony sector, they are required to have a license first. MCMC normally did not allow foreign entrant if they don't have PRI access and non-disclosure agreement with Telekom Malaysia Berhad. Concurrently, when service providers proceed to do non-disclosure agreement with TM, to terminate call through PSTN, ASP license is required to show them to process. Logically, it is necessary to have a license first, as legislation formatted in this way, but it is really impossible to do non-disclosure agreement with TM before getting an ASP license from MCMC. So, this issue can be the first obstacle of starting business for the foreign investors in Malaysia which supports the first questions of this research.

Information gathered from interview where various issues appeared is directly related to research questions and research objectives as well. In the first official meeting conducted by the researcher about the processing of getting license and PRI access at TM where get details information about it along with the role and function of TM wholesale domestic as well as international access regulations and its requirement.

The essential matter of getting PRI and international access is, pre-requisite of bank guarantee to any local Malaysian bank. In this regards, few things need to submit such as; certified copy of NIC of FD depositor, certified copy of board of regulation, bank guarantee form, and letter of award or contract from TM. Due to pre-requisite and obligatory matter of licensing, Application Service Providers have to terminate through TM or any other incumbent operators at Malaysia. Telekom Malaysia offers two types of TMCH termination package (flat rate access/usage base access) where providers can access.

There are few issues appeared while the interview session was going on with manager of TM wholesale carrier, such as;

1. Pricing imbalance with others incumbent operator in the world. Minute charges are higher for VoIP service and service providers can buy easily from overseas carrier cheaply.

2. Volume base price rate is another risky issue for VoIP providers. Due to the strong competition, market entrant or existing companies could not afford with high price minutes charges. High volume Usage Company will get more privilege than low volume usage. In this case, there is risk for the low volume of Usage Company washing out from the market.

3. Quality of service for termination in all countries is not satisfactory.

4. Testing period is sometimes too lengthy though it depends on the schedule.

5. There is very limited follow up approach about the call service, quality of service and redundancy. Rather it mostly focused on rules, regulations and pricing.

6. There is no protection law for the Application Service Provider.

7. Service Providers have to deposit huge amount as a bank guarantee. After confirmation of BG, traffic transmission over TMCH to destination country will allow. So, during interview session, researcher gets to know approximate amount of BG for termination access that is mandatory for ASP. But such type of rules is not following other countries like Japan, Australia, and South Korea. As they don't follow any typical rules which are disfavoring for the providers, so they are becoming the market leader in the world.

However, there are many concepts and ideas had shared during interview and the official meeting, TM officials knows well about their shortcomings of usage base pricing and failure to keep their clients. Furthermore, Application Service Providers are switching into different network service providers to get cheaper price mostly, from Japan, Australia, France, UK, and USA, though officially is not allowed without having individual license. As a result Application Service Providers are not transparent with TM. The researcher found, ASP license holder somehow earning money by using outside carriers volume of minutes as a cheaper and quality of sound, MCMC and TM will never know their original condition of their service. So, they might not take any initiatives to take care of them. But if rules and regulations are favoring the service providers', transparency will be appearing like other EU countries.

While the researcher managed to take interview from licensing department of Malaysian Communications and Multimedia commissions, there are certain matters revealed which are undoubtedly complicated for the ASP license holder. Licensing issue, monitoring, limitation, market entrant flexibility along with pricing issue have identified. MCMC determines licensing regime to foster the market growth and to reach target of the "National Objective". Secondary data proves the purpose of licensing, is to promote fair competition and development the IP telephony market. But the researcher's observation shows, though there are few technical and diplomatic method give the impression during the licensing procedure that, somehow, it appeared "easy to get license but difficult to survive" in the competitive market. The data shows that, licensing areas of four categories of services. In 2007, there were 64 Network Facilities Providers (NFP) and 26 is under class license; Network Service Providers is 69 where 28 is under class license; Content Application Provider (CASP) was 20 in numbers, but the Application Service Providers was 370 though they all are not providing IP telephony services. In 2010 only Application service Providers are 490. So, data shows that, license holder is increasing over the time but the market stability and sustainability is not achieved due to the high degree of regulation and not having protection law for the ASP Service Providers. Unlike UK market strategy, where Ofcom intervene towards the public policy goal to implement the efficient principle of consistent and transparent regulation for market intervention if needed to apply. But in Malaysian regulatory body does not have any principle to protect the market entrant. Ofcom sets the seven principles which helped to growth of the market and policy implementation as well.

Furthermore, the researcher obtained data from IP telephony service providers, mostly from the foreign joint ventures; of how is the business process and rules affect to foster the market growth and their benefits. So, during the attachment of the researcher in IP Telephony Company observed that, the most competitive and strategic market environment which is rapidly expanding due to lower price and less infrastructure cost. It is observed that, few company have the powerful billing capabilities, scalable and cost effective calling card services but failed to survive due to regulations and huge price difference between global pricing with TM price. However, few reasons are being exposed for market loss and switched to other business from the Service Providers point of view;

i. It is well known that due to the less infrastructure cost VoIP pricing is always less than the PSTN pricing. But, if it is being imposed to buy with high price, definitely Application Service Provider will no longer continue and sustain in their VoIP service market.

ii. Very tough to get access in the market place and after incorporating difficult to sustain.

iii. Imbalance competition in this market due to the direct involvement of local network service providers. The assumption of IP telephony companies are that, existing network service operator can get more privileges as they have a strong local market and huge subscriber to get involve into the VoIP arena easily to cover up the market. In a sense, few operators are accessing through direct calling system.

iv. The issue of global access price is another factor for entrant and the IP telephony service providers. It is expected that, price should be similar with local PSTN price though volume usage is another key factor where price is determined by the volume usage.

v. The researcher get to know from few of the IP telephony service providers that, as an incumbent operator TM is the driving force for selecting pricing issue. TM itself doing VoIP service as a competitor of the IP telephony service providers.

vi. Due to the large volume of usage DIGI is able to offer cheap price rate with a form of "promotional offer'' to turned into them and getting the market lead but IP telephony service provider is not able to do so, as they don't have large market and popularity.

vii. In Malaysian IP telephony market, the target group is illiterate working class, student, and immigrant mostly from South Asian region. Among them, illiterate working class is the majority. So, if any promotional offercomes, incumbent operator easily can divert them.

Another issue came into the spotlight when the researcher took interview of one of the IP telephony service provider's CEO that, the restrictions of toll-free number from mobile which affect the market growth. Unlike Japan, Malaysian Authority is not allowing toll-free from mobile which is more convenient to users and easy to maintain for service providers. These all issues are influenced the growth of the IP telephony market. As a result, they are switching in different business or investing another country where plenty of opportunities are giving available for them.

However, the researcher also analyzed the secondary data from the MCMC publications, company records, and industry analysis; annual report, articles and journal, ITU website source along with practical observations where IP telephony status can see, regulatory issues and market information worldwide, pricing challenges, Malaysian legal framework, national policy objectives, licensing restrictions and market fostering framework all are relating to the research objectives and research questions as well.

5.2 Findings/ Issues in Brief

The major issue appears in primary research is pricing with "DIGI's promotional" offer in a mode of direct calling system with cheaper price. This is the first barrier to continue the IP telephony service in Malaysia. The target of DIGI's business is mostly in South Asia. At the same time foreign joint investors are also choses the same region, so the imbalance competition raised as the method of calling is different.

1. Calling card and call shop platform is still in a good position in the market but direct calling method can deter to the IP telephony service growth.
2. PC-to-PC, PC-to-Phone call has a strong demand because of its less call rate compare to PSTN call. But, this service is not convenient for all types of users.
3. To some extend DIGI's way of offering low cost is a new policy to lead VoIP market. So the protection rules can be implemented by MCMC towards anti-monopoly market to foster competition and invite more foreign investors in IP telephone service.
4. Celcom and Tunetalk also became the barrier of IP telephony service. Like DIGI, celcom has a large local market, so that, if they offer any type of service for international long distance calling card and IDD service, clients are interest to receive. Though they offer with low price or same price compare to the IP telephony service providers, clients are divert to them only for direct calling method of the promotional offer though it is not permanent. Call back approach is lengthy process to get connection, first have to make a call to toll-free number, and then it will answer from the system, then have to enter destination number followed by #. Calling card approaches is a bit different then calls back system. But if, people get access to the "direct calling method", certainly they will divert to easier way of use method, because by instinct people are always chose easier method. But after a certain period, DIGI and Celcom have reverted in the previous price system. For e.g. calling to BD through DCampus was only 20 cent per minute but suddenly they reverted in 40 cent after few month, that is higher than IP telephony price. So, the ways incumbent operators enter into the market causes the block for IP telephony service provider's service and drag them in a disastrous situation.

There are few more issues appeared from the secondary data analysis.

5. Regulatory issue can either promoting the market or prohibit to entering the market or even if they enter could make difficulties to survive. Even it can be foster fair and competitive market.
6. TM price is an imbalance with global price. But in order to promote more investment prices

should be efficient and equivalent to global access price.

7. One of the national objectives for Malaysia is become a “global centre and hub for communications”.So, fostering the market growth foreign investment is necessary.
8. The data show the worldwide regulatory issues and challenges as well as differences from one country to another where few countries are gaining advantages and some other declining. Besides, some countries follow market-driven approach which brings positive impact to IP telephony services and its growth.
9. ITU World Telecommunication Policy Forum (WTPF) provides a policy forum, where ITU member states can share and discus the regulatory challenges and emerging telecommunication policy for changing telecommunication environment over time based on their need and demand. So, WTPF helps to know the telecommunication regulation worldwide with appropriate opinions, market structure, and leading force of the IP telephony service which are reflected throughout the paper.

6. Conclusion and Recommendations

From the analyses of IP telephony perspective, it is anticipated that, as rapid revenue generated source, IP telephony becomes the key technology in compare to the Circuit Switched Network to Packet Switched Network and opens up a new opportunity for voice and data communication all over the world. Besides, IP telephony in its way forward is expected because of its less infrastructure cost and quick return on investment. It also found that, this sector is full of competition whoever grabs this opportunity gains the market strength and access to the new area of competition. So, this paper highlighted that, as a new technology, VoIP takes place in an approach of alternative voice calls capable of provide the higher efficiency of services with high level of service commitment. As revenue generating source, there are several countries gaining the extensive market access with a broad geographical coverage.

So, the issues of pricing and regulatory challenges affect the growth of the market and sustainability of IP telephony business environment. The reason of declining IP telephony market has critically analyzed through the practical observation with the issue of “obstacle and Survive”. Accessing into the Malaysian IP telephony market is somehow difficult but after entering it is also very tough to survive due to strong competition among the IP telephony service providers. Furthermore, service providers wet to backward or fall in a disastrous situation in the market by facing challenges of several incumbent operators which have been clearly stated throughout this research.

This research also finds out the limitations of regulatory authorities where no protection rules been imposed. On the other hand, IP telephony service providers are declining because of not having the protection rules but several countries have a protection rules and they apply when needed. This paper also suggests to reforming and adapting new law followed by the other successful countries like Japan, Australia, UK and South Korea. So, ASP could contribute to grow the economy and can reach their expectation to become the “global hub” in the world of communications. In this study, it is significant that, the world leading service providers and operators focus point is to meet the customer’s needs and demand with full of satisfaction and creates the new opportunities worldwide. Besides, this research focused few issues and challenges of IP telephony, business value, market driver and economic aspects.

Apart from these issues, this paper clearly stated the MCMC legislation and its impact upon the service providers. Besides, the role of TM and the issue of PSTN pricing challenges clearly identified. Alongside, the role of ITU also indicates the issue of pricing and regulatory framework worldwide. Finally this paper concluded by giving the few possible solutions to revert the IP telephony market growth and its expansion.

6.1 Recommendations

Due to the low infrastructure deployment cost, IP telephony service is still considered as a driving force and reliable for the voice and data communications whereas revenue generating business opportunities service provider can generate more revenue to contribute the economy. So the well-planned and regulatory reform in favor of service providers will attract more foreign investment to expand the market.

Besides, in order to sustain and growth of the market, it is needed to maintain licensing regime which would be the neutral and easy access along with minimum licensing condition and protect consumer interest to proceed to the fair and competitive market. For this government initiatives would implement which would be updated and similar to other successful countries model.

There is another issue concerned for IP telephony service (providers) that, in order to proceed IP telephony service, providers must be allowed to access to PSTN with lower price, so that they can continue buying from Telekom Malaysia Berhad. The flow of the market would be constant and rising which will determined sustainability of the market and on the way forward to the further development in this sector. So the accurate and low price will increase more investment in this sector and no subsidize will be occur to control over the market from the single incumbent providers. TM should offer new price scheme followed by global price which will stimulate the market demand and will protect from decline of IP telephony service provider's ratio.

Furthermore, VoIP will face more challenges of regulatory issues that directly affected the economic feasibility. If the regulatory authority does not fix price to protect the Application Service Providers from declining especially foreign joint or foreign individual investment market will be unstable.

However, the rise of IP telephony all over the world brings the revolutionary changes of communication but regulatory issues and policies are needed to reform and set principles to meet the economic challenges as this sector offers huge commercial opportunities for the service providers. Regulatory reformation will facilitate the IP telephony market place. Besides the adaptation of the "market-driven" approach will ultimately lead to efficient and fair competitive market place for the IP telephony service providers. There is a hope of reformation of the regulation, foreign investment in this sector will increase and will contribute to the Malaysian economy as well.

Hence, the major issue appears as one of the National Objectives of Malaysia to become a "Global Hub". The suggested idea can possibly be applied if foreign investment as well as building the local infrastructure in this sector will contribute more one to build the national economy. So, market observation policy by the MCMC and TM need to implement to protect them from declining.

References

[1] IP Telephony Industry Report, Suruhanjaya Komunikasi dan Multimedia, Malaysia: SKMM, 2007.

[2] IP Telephony Industry Report, Suruhanjaya Komunikasi dan Multimedia, Malaysia: SKMM, 2007

[3] Rajib, R. Shah, VoIP Challenges and Opportunities, Research and Network Strategy, Texas: 2003.

[4] William A Yarberry Jr, Computer Telephony Integration-2nd Edition, New York: CRC Press, 2003.

[5] IP telephony Industry Report, Suruhanjaya Komunikasi dan Multimedia, Malaysia: SKMM, 2007.

[6] ITU E-Strategies Unit, The Essential Report on IP telephony, ITU: 2003.

[7] ITU E-Strategies Unit, The Essential Report on IP telephony, ITU: 200.

[8] Voice Over IP Telephony, http://voip-facts.net/p2p.php, 2006.

[9] Daniel B. Garrie and Rebeca, Wong, "Privacy in Electronic Communications: The regulation of VoIP in the EU and the United States", Computer Telecommunications Law Review, pp. 139-146, 2009.

[10] Daniel B. Garrie and Rebeca, Wong, "Privacy in Electronic Communications: The regulation of VoIP in the EU and the United States", Computer Telecommunications Law Review, pp. 139-146, 2009

[11] William A Yarberry Jr, Computer Telephony Integration- 2nd Edition, New York: CRC Press, 2003

[12] IP telephony Industry Report, Suruhanjaya Komunikasi dan Multimedia, Malaysia: SKMM, 2007.

[13] J. Harris, IP telephony Shows Strong market Growth: TECH WATCH, March 1, 2010.

[14] Chris, Robert, "Voice Over IP", Centre for Critical Infrastructutr Protection, Welllington, Newzeland, 2005.

[15] www.itxc.com

[16] MIS-ASIA, the home of enterprise IT in Asia, Report: Avaya leads APAC IP telephony, July 4, 2008.

[17] Guideline of the Provisioning of VoIP Service, Malaysian Communications and Multimedia Commission, www.skmm.com.my

[18] http://www.idc.com

[19]James R. Ransom and John R. Ritinghouse, "VoIP Security" Burlington MA 01803, USA, Elsevier Digital Press, 2005

[20] James R. Ransom and John R. Ritinghouse, "VoIP Security Burlington MA 01803, USA, Elsevier Digital Press, 2005

[21] Konrad, L. Trope, Esq, VoIP Deployment and Regulation in Asia, Presentation Paper, California University Tower, Novo Law Group, (2006)

[22] Jonathan Davidson and James Peter, A Systemic Approach to Understanding the Basics of Voice Over IP: Voice Over IP Fundamentals, Indiana, USA: Cisco Press, 2000.

[23] Bejil, de, Paul and Petiz, Access Regulation and the Adoption of VoIP, International University of Germany, 2006.

[24] VocalScape Report, From Business Overview of VoIP in Malaysia, November 2004. http://voip.blogs.com/pbx/2004/11/business_overvi.html

[25] David, Bech and Jonathan Sallet, The Challenges of Classification: Emerging VoIP Regulation in Europe and the United States, Berkeley, USA: 2005.

[26] Paul W.J. and Martin Peitz, Access Regulation and the Adoption of VoIP, Germany: 2006.

[27] Paul W.J. and Martin Peitz, Access Regulation and the Adoption of VoIP, Germany: 2006.

[28] IP telephony Industry Report, Suruhanjaya Komunikasi dan Multimedia, Malaysia: SKMM, 2007.

[29] IP telephony Industry Report, Suruhanjaya Komunikasi dan Multimedia, Malaysia: SKMM, 2007.

[30] Report on World Telecommunication Policy Forum (WTPF), Geneva, 7-9 March, 2001.

[31] World Wide VoIP Regulation and Market Information, Ministry of Commerce, USA.

[32] IP telephony Industry Report, Suruhanjaya Komunikasi dan Multimedia, Malaysia: SKMM, 2007.

[33] IP telephony Industry Report, Suruhanjaya Komunikasi dan Multimedia, Malaysia: SKMM, 2007.

[34]Fukahori,Michiko, Voice Over Internet Protocol. *UNESCAP* United Nations Economic and Social Commission for Asia and the Pacific.

[35] World Wide VoIP Regulation and Market Information, Ministry of Commerce, USA.

[36] The Report of the SG on IP Telephony International

Telecommunication Union, World Telecommunication Policy Forum, Final Report, ITU: 2001.
[37] The Report of the SG on IP Telephony International Telecommunication Union, World Telecommunication Policy Forum, Final Report, ITU: 2001.
[38] Malaysian Communication and Multimedia Commission, selcted fact and figures,Q1, Communication and Mutimedia Act, 1998, Ministry of Engery, water and Communicaitons, Malaysia, 2010. http://www.skmm.gov.my
[39] Malaysian Communication and Multimedia Commission, selcted fact and figures,Q1, Communication and Mutimedia Act, 1998, Ministry of Engery, water and Communicaitons, Malaysia, 2010. http://www.skmm.gov.my
[40] The Essential Report on IP Telephony, Geneva: 2003

http://www.itu.int/en/Pages/default.aspx

[41] Editiors Note, From Japan Registration or Notificaiton, ICT Regulation Toolkit, ITU: 2010
[42] Editors Note International Telecommunication Union, ICT Regulation Toolkit the Joint Production of Infodev and the International Telecommunication Union.
[43] http://www.ictregulationtoolkit.org/en/index.html

[44] Practice Note, The Regulatory Framework for General Authorization, 2010. http://www.ictregulationtoolkit.org/en/Index.html
[45] Practice Note, The Regulatory Framework for General Authorization, 2010. http://www.ictregulationtoolkit.org/en/Index.html
[46] Practice Note, The Regulatory Framework for General Authorization, 2010. http://www.ictregulationtoolkit.org/en/Index.html
[47] http://www.ofcom.org.uk/
[48] Raslan, Loong, Malaysia Communication Policy: Getting the Deal Through- Telecoms and Media, 2006.
[49] ITU Report.
50] Telekom Malaysia from TM Products and Services, VoIP Premium Termination, 2010
[51] Telekom Malaysia from TM Products and Services, VoIP Premium Termination, 2010
[52] TM Products and Services, VoIP Termination, How it Works, July 2010.
[53]TM Products and Services, Termination Overview, July 2010.
[54] http://www.tm.com.my/sme/products/Pages/Home.aspx
[55] http://www.itu.int/en/Pages/default.aspx
[56] http://www.itu.int/en/Pages/default.aspx
[57] http://www.itu.int/en/Pages/default.aspx
[58] SKMM website
[58] http://www.itu.int/en/Pages/default.aspx

First Author The only author of this paper is Md. Khaled Shukran a student of International Masters in Information Management at Asia Europe Institute, University of Malaya, Kuala Lumpur, Malaysia where he has completed his Masters with full scholarship. He has also completed B.A. Honors in English Language and Literature from Northern University Bangladesh. This author has a couple of working experiences in different institutions both in Bangladesh and Malaysia. He is a very hard working, active and self-motivated person. His intention is to be an academician. His article is also going to publish at IEEE very soon. His major concentrating areas are, Knowledge economy, Knowledge management, ICT, IS, IP Telephony and others related field.

High Throughput and Low Power NoC

Magdy El-Moursy[1], *Member IEEE* and Mohamed Abdelgany[2]

**[1] Mentor Graphics Corporation
Cairo, Egypt**

**[2] Electronics Research Institute
Cairo, Egypt
German University in Cairo, Cairo, Egypt**

Abstract

The High throughput architecture to achieve high performance Networks-on-Chip (NoC) is proposed. The throughput is increased by more than 38% while preserving the average latency. The area of the network switch is decreased by 18%. The required metal resources for the proposed architecture are increased by less than 10% as compared to the required metal resources for the conventional NoC architecture. Power characteristics of different high throughput NoC architectures are developed. The extra power dissipation of the proposed high throughput NoC is as low as 1% of the total power dissipation. Among different NoC topologies, High Throughput Butter Fat Tree (HTBFT) requires the minimum extra power dissipation and metal resources.

Keywords: *Network-on-Chip, Throughput, Power Dissipation, Topology.*

1. Introduction

As the number and functionality of intellectual property blocks (IPs) in System on Chips (SoCs) increase, complexity of interconnection architectures of the SoCs have also been increased. Different research articles have been published in high performance SoCs. However, the system scalability and bandwidth are limited. As described in [1]-[5], NoCs are emerging as the best replacement for the existing interconnection architectures. Many NoC topologies have been proposed in the past, e.g., CLICHÉ [1], SPIN [2], Octagon [3] and Butterfly Fat Tree [4]. Different research articles in architectural and conceptual aspects of NoC such as, topology selection, quality of service (QoS) [5], design automation [6], performance evaluation [7], and verification have been reported. NoCs provide different set of constraints in the design paradigm. High throughput and low latency are the desirable characteristics of a multi processing system.

Previous articles have taken a top down approach (a high level analysis of NoC) and they did not touch the issues on a circuit level. However, little research has been reported on the circuit design issues [8]. Although they were implemented and verified on silicon, they were only focusing on implementing limited set of topologies. In large scale NoCs, power dissipation should be minimized for cost efficient implementation. Many papers have been published in NoCs. They were only focusing on performance and scalability issues rather than power efficiency. Scaling with power reduction is the trend in future technologies. Application specific techniques are required to reduce power dissipation of NoCs.

The main focus of this paper is to present a high throughput interconnect architecture for network on chip. The circuit implementation issues are considered in the proposed architecture. The switch structure along with the interconnect architecture are shown in Figure 1 for 2 IPs and 2 switches. The proposed architecture is applied to different NoCs topologies. Low power switch is also proposed to achieve power-efficient NoC. The efficiency and performance are evaluated.

To the best of our knowledge, this is the first in depth analysis on circuit level to optimize performance of different NoC topologies. The paper is organized as follows: In Section 2, the proposed port architecture is presented. The new High Throughput architecture is described in section 3. In Section 4, closed form expressions for the power dissipation in different high throughput architectures are developed. The performance improvement and circuit overhead of the proposed architecture are provided in Section 5. Finally, conclusions are summarized in section 6.

2. Port Architecture

Each port of the switch includes input virtual channels, output virtual channels, header decoder, controller, input arbiter and output arbiter as shown in [4]. The input arbiter consists of a priority matrix and grant circuit. The priority matrix stores the priorities of the requests. A dedicated circuit generates the grant signals to allow only one virtual channel to access a physical port. The messages are divided into fixed length flow control units (flits). When the granted virtual channel stores one whole flit, it sends a full signal to controller. If it is a header flit, the header decoder determines the destination. The controller checks the status of destination port. If it is available, the path between input and output is established. The flits from more than one input port may simultaneously try to access a particular output port. The output arbiter is used to allow only one input port to access an output port. Virtual channels consist of several buffers controlled by a multiplexer and an arbiter which grants access for only one virtual channel at a time according to the request priority. Once the request succeeds, its priority is set to be the lowest among all other requests.

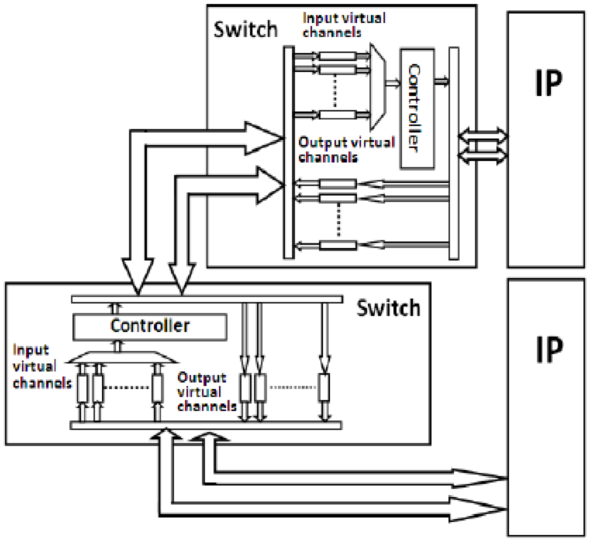
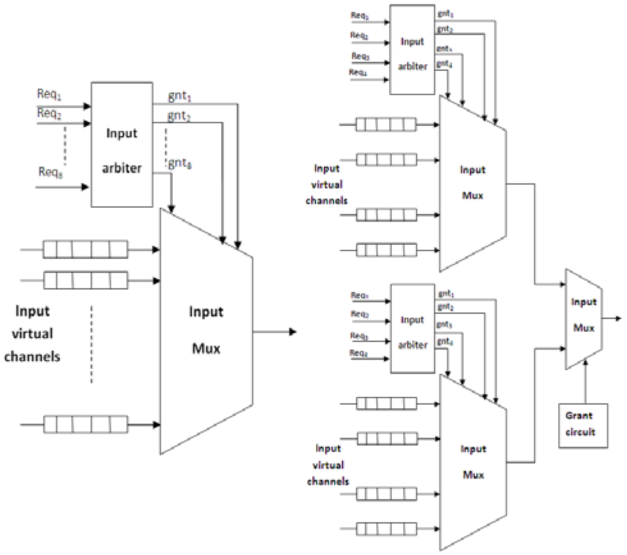


Fig. 1: Proposed high throughput architecture.

In the proposed architecture, rather than using one multiplexer and one arbiter to control the virtual channels, two multiplexers and two arbiters are employed as shown in Figure 2. Using the proposed technique, the virtual channels are divided into two groups; each group is controlled by one multiplexer and one arbiter. Each group of virtual channels is supported by one interconnect bus as described in section 3. However trivial it may look, the proposed port architecture has a great influence on the switch frequency and the throughput of the network. Let us consider an example with 8 virtual channels. In the NoC architecture, 8x8 input arbiter and 8x1 multiplexer are needed to control the input virtual channels as shown in Figure 2 (a). The 8x8 input arbiter consists of 8x8 grant circuit and 8x8 priority matrix. In the proposed architecture, two 4x4 input arbiters, two 4x1 multiplexers, 2x1 multiplexers and 2x2 grant circuit are integrated to allow only one virtual channel to access a physical port as shown in Figure 2 (b). The 4x4 input arbiter consists of 4x4 grant circuit and 4x4 priority matrix. The values of the grant signals are determined by the priority matrix. The number of grant signals equals the number of requests and the number of selection signals of the multiplexer. The area of two 4x4 input arbiters is smaller than the area of 8x8 input arbiter. Also, the area of two 4x1 multiplexers is smaller than the area of 8x1multiplexer. Consequently, the required area to implement the proposed switch with the proposed architecture is less than the required area to implement the conventional switch.

In order to divide a 4x1 multiplexer into three 2x1 multiplexers, the 4x4 input arbiter should be divided into three 2x2 input arbiters. The grant signals which are generated by three 2x2 input arbiter (6 signals) are not the same grant signals generated by the 4x4 input arbiter (4 signals). Therefore, the 4x4 input arbiter can not be replaced by three 2x2 input arbiters unless the number of interconnect buses is increased to be equal to the number of virtual channels groups. Therefore, the proposed architecture in Figure 2 (b) is the optimum to allow eight virtual channels in the port. By increasing the number of interconnects, the metal resources and power dissipation are increased as described in Section 6.



(a) (b)

Fig. 2 (a) Circuit diagram of switch port, (b) circuit diagram of High Throughput switch port.

Without circuit optimization, the change in the maximum frequency of the switch with the number of virtual channels in the conventional BFT switch is shown in Figure 3. When the number of virtual channels is increased beyond four, the maximum frequency of the switch is decreased for BFT architecture. Throughput is a parameter that measures the rate in which message traffic can be sent across a communication network. It is defined by [7]:

$$TP = \frac{(\text{number of messages completed}) * (\text{message length})}{(\text{number of IP blocks}) * (\text{total time})} \quad (1)$$

The throughput is proportional to the number of completed messages. The number of completed messages increases with the number of virtual channels. Total transfer time of messages decreases with the increase in frequency of the switch. Therefore the throughput can be improved by increasing the number of virtual channels or by increasing the operating frequency of the switch. The throughput is saturated when the number of virtual channels is increased beyond four [7]. On the other hand, the average message latency increases with the number of virtual channels. To keep the latency low while preserving the throughput, only four virtual channels are used in [7].

The proposed High Throughput BFT (HTBFT) switch is smaller than the BFT switch. Therefore, the maximum frequency of the switch is improved. The change in the maximum frequency of the proposed switch with the number of virtual channels is shown in Figure 3 for HT-BFT architecture. With the proposed switch architecture, the number of virtual channels could be increased up to eight without significant reduction in the operating frequency. The frequency of the network switch is characterized for different network topologies using the proposed architecture as shown in Figure 4. As compared to the conventional architecture, the operating frequency of the proposed architectures is decreased when the number of virtual channels is higher than eight rather than four in the conventional architecture. Doubling the number of virtual channels does not degrade the frequency of the switch (rather than 4 virtual channels, 8 virtual channels could be used in the proposed architecture). However, a severe increase in the number of virtual channels (more than 8) could degrade performance.

Increasing the number of virtual channels would increase the traffic going through the links (interconnects) between the switches, increasing the contention on the bus and increasing the latency that each flit experiences. In order to improve throughput, the links (interconnects) connecting the switches with each other should be increased. Since the number of virtual channels could be doubled (from four in the conventional architecture to eight in the proposed architecture), doubling the number of virtual channels between switches is proposed.

Let us consider an example of BFT architecture. The HTBFT architecture decreases the area of switch by 18%. Consequently, a system with eight virtual channels achieves high throughput, high frequency and low latency while the area of the design is optimized. The architecture of different NoC topologies to achieve high throughput network is described in section 3.

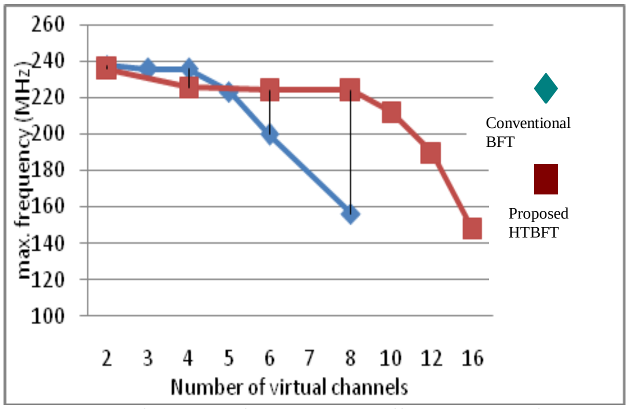


Fig. 3 Maximum frequency of a switch with different number of virtual channels for conventional BFT and proposed HTBFT.

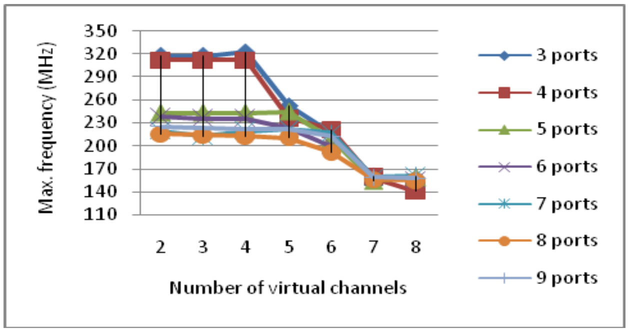


Fig. 4 Maximum frequency of a switch with different number of virtual channels for different NoC topologies of the proposed architecture.

3. High Throughput Architecture

To A novel interconnect template to integrate IP blocks in NoC is proposed. In the proposed architecture, rather than using a single interconnect bus between each two elements of NoC (IP block and switch or two switches), two buses are employed. The number of virtual channels can be doubled to get higher throughput. To maintain the average latency, each bus supports half the number of virtual channels. Increasing the number of buses between two switches could improve the throughput by optimizing the design of the switch on the circuit level as shown in Section II. However, using two buses to connect two switches implies using more metal resources and may be silicon area for the repeaters within the long interconnects. The overhead of the proposed architecture is discussed in Section 5.

A novel interconnect template to integrate IP blocks using High Throughput Butter Fly Fat Tree (HTBFT) architecture is proposed. Each group of 4 IPs (no. 0, no. 1, no.2 and no.3) in Figure 5 needs one switch (no.4). Each switch in the first level (no. 4) connects to each switch in the second level (no. 5) by 2 buses. Each bus supports half the number of virtual channels. Therefore, the throughput can be improved while preserving the average latency.

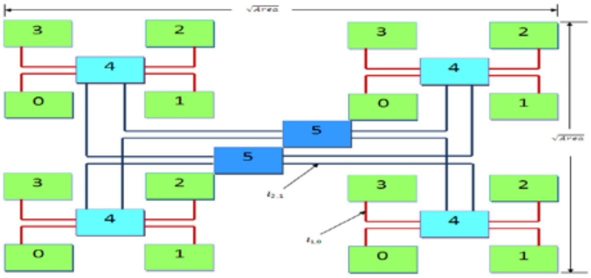


Fig. 5 Interconnect architecture of HTBFT

The interconnect template to integrate IP blocks using High Throughput architecture is implemented to CLICHÉ (to become High Throughput CLICHÉ, HTCLICHÉ), Octagon (to become High Throughput Octagon, HT-Octagon), SPIN (to become High Throughput SPIN, HTSPIN) architectures, in which double the number of interconnects is needed. The throughput improvement is presented in section V for each topology.

Power estimation is very important aspect of NoC design. The average power dissipation of NoC port is obtained. The switch is implemented on the transistor level using ASIC design flow. For different NoC topologies, the average power dissipation of the switch is determined. Closed form expressions are developed for each topology in section 4.

4. Power Characteristics

To Communication network on chip contains three primary components; network switch, interswitch links (interconnects), and repeaters within interswitch links. Including different sources of power dissipation in NoC, the total power dissipation of on chip network is defined as follows:

$$P_{total} = P_{switches} + P_{interconnect} + P_{reps}\ , \qquad (2)$$

$$P_{switches} = P_{switching} + P_{leakage}\ , \qquad (3)$$

where P_{total} is the total power dissipation of the network. $P_{switches}$ is the power dissipation in the switches. $P_{interconnect}$ is the total power dissipation of interswitch links. P_{reps} is the total power dissipation of the repeaters which are required for long interconnects. $P_{switching}$ and $P_{leakage}$ are the switching and leakage power of the switch, respectively. The number of repeaters depends on the length of the interswich link. According to the topology of NoC interconnects, the interswitch wire lengths, the number of repeaters and the number of switches can be determined a priori.

$$P_{interconnect} = cV_{dd}^{\ 2} f, \qquad (4)$$

$$P_{reps} = P_{reps\text{-}dyn} + P_{reps\text{-}SC} + P_{reps\text{-}leakage}, \qquad (5)$$

$$P_{reps\text{-}dyn} = N_{rep}H_{opt}C_0V_{dd}^{\ 2} f, \qquad (6)$$

where $P_{reps\text{-}dyn}$ is the total dynamic power dissipation of repeaters, N_{rep} is the number of repeaters, H_{opt} is the optimum repeater size, C_0 is the input capacitance of a minimum size repeater, V_{dd} is the supply voltage and f is the switching frequency. $P_{reps\text{-}SC}$ is the total short-circuit power of the repeaters. $P_{reps\text{-}leakage}$ is the total leakage power dissipation of the repeaters. c is the interswitch link capacitance. Closed form expressions for the power dissipation of different high throughput NoC architectures are described in the following subsections.

4.1 High Throughput Butterfly Fat Tree

In the HTBFT, the interconnection is performed on levels of switching. The number of switching levels can be expressed as $log_2N - 3$, where N is the number of IP blocks. The total number of switches in the first level is $N/4$. At each subsequent level, the number of required switches reduces by a factor of 2 as shown in Figure 5. The interswitch wire length and total number of switches are given by the following expressions:

$$l_{a+1,a} = \frac{\sqrt{Area}}{2^{levels-a}}, \qquad (7)$$

$$N_{switches-HTBFT} = \frac{N}{4}\left(\frac{1-(1/2)^{levels}}{1-1/2}\right), \qquad (8)$$

where $l_{\ a+1,a}$ is the length of the wire spanning the distance between level a and level $a + 1$, where a can take integer value between 0 and ($levels$ - 1). In the HTBFT, the total length of interconnects and the total number of repeaters can be determined from the following equations:

$$l_{tot-HTBFT} = \frac{\sqrt{Area}}{2^{(\log_2 N-3)}} NX(levels) X 2N_{wires}, \qquad (9)$$

$$N_{rep-HTBFT} = 2NN_{wires}\left(\left\lfloor\frac{l_{1,0}}{K_{opt}}\right\rfloor + \ldots \frac{1}{2^{N-1}}\left\lfloor\frac{l_{lev,lev-1}}{K_{opt}}\right\rfloor\right), \qquad (10)$$

where K_{opt} is the optimum length of the global interconnect [9]. Using the number of switches, the total length of interconnects and the total number of repeaters, the total power dissipation of HTBFT architecture ($P_{tot\text{-}HTBFT}$) is determined.

$$P_{tot-HTBFT} = 3\frac{N}{2}\left(\frac{1-(1/2)^{levels}}{1-1/2}\right)P_{port} +$$
$$\frac{\sqrt{Area}}{2^{(\log_2 N-3)}} NX(\log_2 N-3)X\,2N_{wires}cV_{dd}{}^2 f +$$
$$2NN_{wires}\left(\left\lfloor\frac{l_{1,0}}{K_{opt}}\right\rfloor + \ldots \frac{1}{2^{N-1}}\left\lfloor\frac{l_{lev,lev-1}}{K_{opt}}\right\rfloor\right)H_{opt}C_0V_{dd}{}^2 f. \qquad (11)$$

4.2 High Throughput CLICHÉ

In HTCLICHÉ, the number of switches equals the number of IPs. The interswitch wire lengths can be determined from the following expression:

$$l_{HTCLICHE} = \frac{\sqrt{Area}}{\sqrt{N}}, \qquad (12)$$

The number of horizontal interswitch wires between switches equals $2\sqrt{N}(\sqrt{N}-1)$. According to the technology node, the optimum length of global interconnects can be obtained. Therefore, the total length of interconnects and the number of repeaters can be calculated by:

$$l_{tot-HTCLICHE} = 4\sqrt{Area}\left(\sqrt{N}-1\right)N_{wires}, \qquad (13)$$
$$N_{rep-TCLICHE} = 4\left\lfloor\frac{\sqrt{Area}}{\sqrt{N}K_{opt}}\right\rfloor\sqrt{N}\left(\sqrt{N}-1\right)N_{wires}, \qquad (14)$$

Using the number of ports, number of switches, total length of interconnects and number of repeaters, the total power dissipation of the HTCLICHÉ architecture can be determined.

$$P_{tot-HTCLICHE} = 5NP_{port} +$$
$$4\left\lfloor\frac{\sqrt{Area}}{\sqrt{N}K_{opt}}\right\rfloor\sqrt{N}\left(\sqrt{N}-1\right)N_{wires}H_{opt}C_0V_{dd}{}^2 f +$$
$$4\sqrt{Area}\left(\sqrt{N}-1\right)N_{wires}cV_{dd}{}^2 f. \qquad (15)$$

4.3 High Throughput Octagon

For HTOctagon, there are four types of interswitch wire lengths: First [wires which connect nodes (1,5) and (4,8)], second [wires which connect nodes (2,6) and (3,7)], third [wires which connect nodes (1,8) and (4,5)], forth [wires which connect nodes (1,2), (2,3), (3,4), (5,6), (6,7) and (7,-8)]. The interswitch wire lengths can be defined by (l_1=3L/4, l_2=13w_l N_{wires} +L/4, l_3=13L/4, l_4=L/4), where L is the length of four nodes which equals $\left(4*\sqrt{\frac{Area}{N}}\right)$. w_l is the summation of the global interconnect width and space. Considering the interswitch wire lengths and the optimum length of global interconnect, the total length of interconnects and number of repeaters can be obtained by:

$$l_{tot-HTOctagon} = (7L+104w_lN_{wires})N_{wires}N_{oct-unit}, \qquad (16)$$
$$N_{rep-HTOctagon} = (4\left\lfloor\frac{3L/4}{K_{opt}}\right\rfloor + 4\left\lfloor\frac{13w_lN_{wires}+L/4}{K_{opt}}\right\rfloor +$$
$$4\left\lfloor\frac{13w_lN_{wires}}{K_{opt}}\right\rfloor + 12\left\lfloor\frac{L/4}{K_{opt}}\right\rfloor)N_{wires}N_{oct-unit}, \qquad (17)$$

$N_{oct-unit}$ is the number of basic octagon unit. The total power dissipation of the HTOctagon architecture is obtained by (22).

$$P_{tot-HTOctagon} = 3NP_{port} +$$
$$(28\sqrt{\frac{Area}{N}} + 104w_lN_{wires})N_{wires}N_{oct-unit}c\,V_{dd}{}^2 f$$
$$+(4\left\lfloor\frac{3L/4}{K_{opt}}\right\rfloor + 4\left\lfloor\frac{13w_lN_{wires}+L/4}{K_{opt}}\right\rfloor$$
$$+4\left\lfloor\frac{13w_lN_{wires}}{K_{opt}}\right\rfloor$$
$$+12\left\lfloor\frac{L/4}{K_{opt}}\right\rfloor)N_{wires}N_{oct-unit}H_{opt}C_0V_{dd}{}^2 f. \qquad (18)$$

4.4 High Throughput SPIN

An interconnect template to integrate IP blocks using HTSPIN architecture was proposed. In large HTSPIN, the total number of switches is *3N/4*. The interswitch wire length can be determined using (7). In HTSPIN, the total length of interconnects and the number of repeaters is defined by:

$$l_{tot-HTSPIN} = 1.75\sqrt{Area}N_{wires}N, \qquad (19)$$
$$N_{rep-HTSPIN} = (\left\lfloor\frac{\sqrt{Area}}{8K_{opt}}\right\rfloor + \left\lfloor\frac{\sqrt{Area}}{4K_{opt}}\right\rfloor$$
$$+\left\lfloor\frac{\sqrt{Area}}{2K_{opt}}\right\rfloor)2N_{wires}N. \qquad (20)$$

The total power dissipation of the HTSPIN architecture can be determined by

$$P_{tot-HTSPIN} = 6NP_{port} + 1.75\sqrt{Area}\,N_{wires}NcV_{dd}^{2}f + \left(\left\lfloor \frac{\sqrt{Area}}{8K_{opt}} \right\rfloor + \left\lfloor \frac{\sqrt{Area}}{4K_{opt}} \right\rfloor + \left\lfloor \frac{\sqrt{Area}}{2K_{opt}} \right\rfloor\right)2N_{wires}NH_{opt}C_{0}V_{dd}^{2}f. \tag{21}$$

4.5 Power Dissipation for Different High Throughput NoC Architectures

According to (11), (15), (18) and (21), the total power dissipation of the network can be expressed as a function of the number of IP blocks. The change in the power dissipation with the number of IP blocks for different high throughput network architectures is shown in Figure 6. The power dissipation for different NoC topologies increases by different rates as the number of IP blocks increases. The HTSPIN and HTOctagon architectures have much higher rate of power dissipation increase. The HTBFT architecture consumes the minimum power as compared to other NoC topologies making HTBFT more attractive as a power efficient NoC topology.

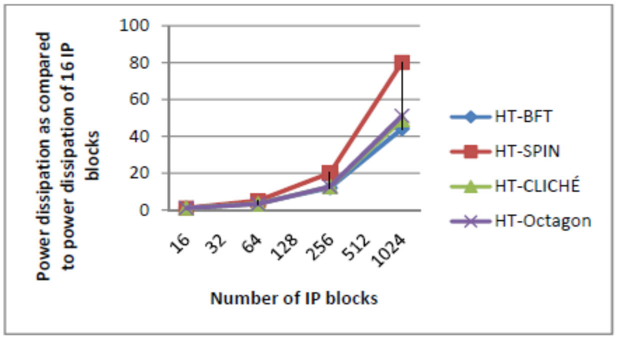


Fig. 6 power dissipation of different NoC topologies

The ratio of the power dissipation in the interswitch links and repeaters as compared to the total power dissipation is shown in Figure 7. For the HTSPIN network, the power dissipation of the interswitch links and repeaters represents 40% of the total power dissipation of the network. For the HTBFT, HTCLICH and HTOctagon, the percent of power dissipation of the interswitch links and repeaters decreases with increasing the number of IP blocks. For future SoC, reducing power dissipation should be focusing on reducing the power of the switches. More detailed results using real example of an SoC are provided in Section 5.

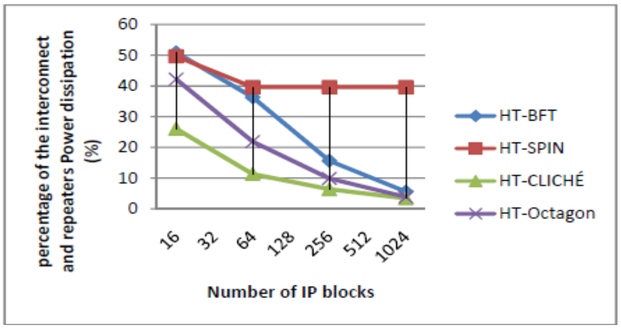


Fig. 7 Power dissipation of interswitch links and repeaters for different NoC architectures.

5. Performance and Overhead Analysis

The proposed high throughput architectures are implemented using Application Specific Integrated Circuit (ASIC) design flow (Leonardo Spectrum synthesis tool), with 90nm technology. Under uniform traffic assumption, the throughput for different NoC architectures is calculated. In the following subsections, the throughput and power dissipation are presented.

5.1 Improvement of the Throughput

The proposed high throughput architecture doubles the number of virtual channels to increase the throughput while preserving the average latency. Therefore, the average latency of HTBFT with 8 virtual channels equals the average latency of BFT with 4 virtual channels. Uniform traffic and maximum operating frequency are assumed to determine the throughput of HTBFT. The change in the throughput with the number of virtual channels for HTBFT and BFT is shown in Figure 8. In the proposed architecture, when the number of virtual channels is increased beyond eight, the throughput saturates. The architecture increases the throughput of the network by 38%. The increase in the throughput for different architectures is presented in Table 1. The maximum improvement is achieved in HTCLICHÉ. The increase in the throughput for HTSPIN is the minimum as compared to the other high throughput architectures.

5.2 Overhead of High Throughput Architecture

With the advance in technology, the number of metal layers increases every generation. Considering a chip size of 20 mm x 20 mm, technology node of 90 nm, and a system of 256 IP blocks, the length of interswitch links for different NoC topologies is obtained. Given the optimum global interconnect width W_{opt} of 935 nm, optimum global interconnect spacing S_{opt} of 477 nm [9], the global interconnect pitch is 1.412 µm (W_{opt} + S_{opt}). Accordingly, the number of global interconnects N_{gi} per layer equals $\frac{\sqrt{Area}}{W_{opt}+S_{opt}}$.

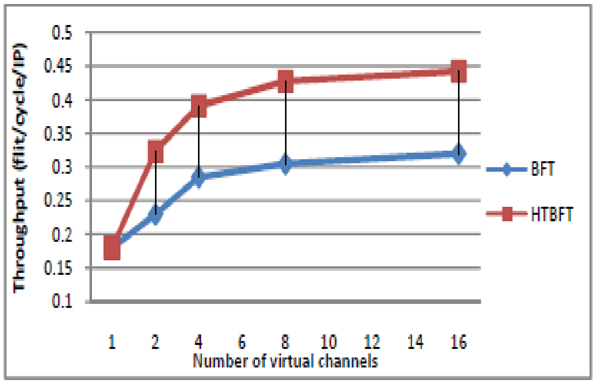


Fig. 8 Throughput for different number of virtual channels

Table 1: The percentage of increase in the throughput for different high throughput architectures

Architecture	Increase in throughput (%)
HT-BFT	38
HT-CLICHÉ	40
HT-Octagon	17
HT-SPIN	12

Using the critical interconnect length of the target technology as 2.54 mm and the optimum repeater size as 174 [9], the number of repeaters is determined. The butterfly fat tree can be laid out in *O(N)* active area (IPs and switches) and *O(log(N))* wiring layers [10]. The basic strategy for wiring is to distribute tree layers in pair of wire layers; one for horizontal wiring $H_{a+1,a}$ and one for vertical wiring $V_{a+1,a}$. The length of horizontal part $H_{a+1,a}$ equals the length of vertical part $V_{a+1,a}$ given that the chip is squared. More than one tree layer can share the same wiring trace.

High throughput architecture has the same number of switches, but the number of wires and repeaters is doubled. The length of interswitch interconnects depend on the number of levels, which depends on the system size. In the circuit implementation of HTBFT, a bus between each two switches has 12 wires, 8 for data and 4 for control signals. Considering a system of 256 IP blocks, the length of $H_{a+1,a}$ and $V_{a+1,a}$ are calculated. The number of wiring levels is seven. The number of repeaters equals 960. The area of the repeaters equals 20880 µm^2 (it is double the area of the repeaters in the conventional BFT). The power dissipation is presented in Table 2. The power dissipation is increased by 6%.

Table 2: Power dissipation of repeaters and switches for HT-/BFT

Architecture	Number of repeaters	Power dissipation in interswitch links (%)	Power reduction (%)
BFT	960	8.5	
HT-BFT	1920	16.2	6

The horizontal wiring is distributed in the metal layer no. 11 and the vertical wiring is distributed in the metal layer no. 12. The total length of horizontal wires equals 4800 mm (it is 5% of the total metal resources available in metal 11). Similarly, the total length of vertical wires is 5% of the total metal resources available in metal 12. For the proposed design, double the number of interswitch links is required to achieve the communication between each two switches. Therefore, the total metal resource to implement the proposed architecture is 10%. The extra metal resources to achieve the proposed architecture are negligible as compared to the available metal resources.

Considering the same die size of 20mm x 20mm and the system size of 256 IPs, the power dissipation and the required metal resources of other NoC topologies are shown in Table 3. Since the interswitch links is short enough in CLICHÉ, there is no need for repeaters within the interconnects. By applying the proposed high throughput architecture, the HTBFT topology requires the minimum area and power dissipation as compared to the other NoC topologies.

Table 3: Power dissipation and metal resources for different NoC

Architecture	Number of repeaters	Power dissipation of interswitch links and repeaters (%)	Metal resources (%)
CLICHÉ	0	5.4	7
HT-CLICHÉ	0	10.5	14
Octagon	3810	5.2	8
HT-Octagon	7680	10.2	16
SPIN	12288	24.8	28
HT-SPIN	24576	40.4	56

As feature size decreases, more IPs could be integrated in a single chip. System overhead is determined for the adopted architectures for different technology nodes as shown in Table 4. The extra power dissipation is 1% of the total power dissipation of the BFT architecture for 45 nm.

With the advance in technology, the available metal resources in the same die size increases. The number of switches is also increased. The required metal resources to implement the HTBFT are increased by smaller rate than the rate of increase of the available metal resources with the advance in technology. The extra metal resources and power dissipation to implement the HTBFT decrease. The extra metal resource for HTBFT is 3% of the available metal resources. The HTBFT is becoming more efficient as technology advances.

Table 4: Power dissipation of interswitch links and repeaters for different technology nodes

Technology node	**Number of IPs**	**Power dissipation of interswitch links and repeaters (%)**			
		HT-BFT	HT-CLICHÉ	HT-Octagon	HT-SPIN
130 nm	361	17.1	14.1	14.1	58.1
90 nm	729	8.3	9.1	9.1	50.3
65 nm	1849	4.6	5.2	5.4	49.1
45nm	5625	1.2	2.7	3.1	43.8

For SPIN, the extra power dissipation to achieve the proposed HTSPIN architecture is 22% of the total power dissipation. The extra metal resources are more than 100% of the available metal resources (metal 11 and metal 12). Two more metal layers are needed to layout the proposed architecture. Therefore, the overhead in the HTSPIN is high. Applying the high throughput architecture on the SPIN topology is not recommended.

However the proposed architecture has an overhead in power dissipation and metal resources, the overhead decreases as technology advances. The proposed architecture is efficient in improving the network throughput. In the future technologies, the proposed architecture is becoming more power efficient as well as throughput efficient. In the following section, an efficient power reduction technique is proposed to make the proposed architecture further efficient from the power dissipation point of view.

6. Conclusions

In this paper, high throughput NoC architecture is proposed. The proposed architecture is applied to different NoC topologies. The area of the switch is decreased by 18% as compared to the area of conventional NoC switch. The total metal resources to implement the proposed high throughput NoC is increased by less than 10%. It is shown that optimizing the circuit can increase the number of virtual channels without degrading the frequency. The throughput of different NoC topologies is improved with the proposed architecture. Throughput is increased by up to 40%.

The power characteristics of different high throughput NoC topologies are presented. The extra power dissipation to achieve the proposed high throughput architecture is as low as 1% of the total power dissipation of the network. The power dissipation of NoC switches is more than 60% of the total power dissipation of the on chip network. The percent of power dissipation of the interswitch links and repeaters decreases with increasing the number of IP blocks. Reducing power dissipation should be focused on reducing the power dissipation of the switches. The proposed switch and network architecture are becoming more efficient as technology advances. Power overhead decreases with the future technologies.

Acknowledgments

The authors would like to thank Prof. Dr. Mohamed Ismail for his advices and directions.

References

[1] S. Kumar et al., "A Network on Chip Architecture and Design Methodology," *The Proc. of the IEEE Computer Society Annual Symposium on VLSI*, Apr. 2002, pp. 117-124.

[2] P. Guerrier and A. Greiner, "A Generic Architecture for On--Chip Packet Switched Interconnections," *The Proc. of Design, Automation and Test in Europe Conference and Exhibition*, Mar. 2000, pp. 250-256.

[3] F. Karim, A. Nguyen, and Sujit Dey, "An Interconnect Architecture for Networking Systems on Chips," *IEEE Micro*, vol.22, no.5, Sep. 2002, pp. 36-45.

[4] P.P. Pande, C. Grecu, A. Ivanov, and R. Saleh, "Design of a Switch for Network on Chip Applications," *The Proc. of The 2003 International Symposium on Circuits and Systems*, vol.5, May 2003, pp. 217220.

[5] E. Bolotin, I. Cidon, R. Ginosar and A. Kolodny, "QNoC: QoS Architecture and Design Process for Network on Chip," *Journal of Systems Architecture*, vol.50, no.23, Feb. 2004, pp. 105-128.

[6] D. Bertozzi, A. Jalabert and S. Murali et al., "NoC Synthesis Flow for Customized Domain Specific Multiprocessor Systems On Chip," *IEEE Transactions on Parallel and Distributed Systems*, vol. 16, no. 2, Feb. 2005, pp. 113-129.

[7] P. P. Pande, C. Grecu, M. Jones, A. Lvanov, and R. Saleh, "Performance Evaluation and Design Trade Offs for Network on –Chip Interconnect Architectures," *IEEE Transaction on Computers*, vol. 54, no. 8, Aug. 2005, pp. 1025-1040.

[8] K. Lee, S.J. Lee, and H.J. Yoo, "Low Power Networks on -Chip for High Performance SoC Design," *IEEE Transactions on Very Large Scale Integration Systems*, vol. 14, no. 2, Feb. 2006, pp.148-160.

[9] X.C. Li, J.F. Mao, H.F. Huang, and Y. Liu, "Global Interconnect Width and Spacing Optimization for Latency, Bandwidth and Power Dissipation," *IEEE Transactions on Electron Devices*, vol. 52, no. 10, Oct. 2005, pp. 2272-2279.

[10] A. Dehon, "Compact, Multilayer Layout for Butterfly Fat Tree," *The Proc. of The ACM Symposium on Parallel algorithm Architectures*, Jul. 2000, pp. 206-215.

Magdy A. El-Moursy was born in Cairo, Egypt in 1974. He received the B.S. degree in electronics and communications engineering (with honors) and the Master's degree in computer networks from Cairo University, Cairo, Egypt, in 1996 and 2000, respectively, and the Master's and the Ph.D. degrees in electrical engineering in the area of high-performance VLSI/IC design from University of Rochester, Rochester, NY, USA, in 2002 and 2004, respectively. In summer of 2003, he was with STMicroelectronics, Advanced System Technology, San Diego, CA, USA. Between

September 2004 and September 2006 he was a Senior Design Engineer at Portland Technology Development, Intel Corporation, Hillsboro, OR, USA. During September 2006 and February 2008 he was assistant professor in the Information Engineering and Technology Department of the German University in Cairo (GUC), Cairo, Egypt. Dr. El-Moursy is currently a Technical Lead in the Mentor Graphics Corporation, Cairo, Egypt. His research interest is in Networks-on-Chip, interconnect design and related circuit level issues in high performance VLSI circuits, clock distribution network design, and low power design. He is the author of more than 30 papers, four book chapters, and one book in the fields of high speed and low power CMOS design techniques and high speed interconnect.

Mohamed Abdelgany was teaching assistant in the Information Engineering and Technology Department of the German University in Cairo (GUC), Cairo, Egypt. He got his Ph.D. in 2009. His research interest is in Networks-on-Chip. He has many papers in the field.

Adverse Conditions and ASR Techniques for Robust Speech User Interface

Urmila Shrawankar[1], Vilas Thakare[2]

[1]IEEE Student Member, Research Scholar, G H Raisoni College of Engg. Nagpur, INDIA

[2] SGB Amravati University, Amravati, INDIA

Abstract

The main motivation for Automatic Speech Recognition (ASR) is efficient interfaces to computers, and for the interfaces to be natural and truly useful, it should provide coverage for a large group of users. The purpose of these tasks is to further improve man-machine communication.

ASR systems exhibit unacceptable degradations in performance when the acoustical environments used for training and testing the system are not the same.

The goal of this research is to increase the robustness of the speech recognition systems with respect to changes in the environment. A system can be labeled as environment-independent if the recognition accuracy for a new environment is the same or higher than that obtained when the system is retrained for that environment. Attaining such performance is the dream of the researchers.

This paper elaborates some of the difficulties with Automatic Speech Recognition (ASR). These difficulties are classified into Speakers characteristics and environmental conditions, and tried to suggest some techniques to compensate variations in speech signal.

This paper focuses on the robustness with respect to speakers' variations and changes in the acoustical environment.

We discussed several different external factors that change the environment and physiological differences that affect the performance of a speech recognition system followed by techniques that are helpful to design a robust ASR system.

Keywords: *Human Factors, Prosody parameters, Environment Conditions, Environment parameters, environment-independent ASR, ASR Techniques.*

1. Introduction

Speech recognition, is commonly known as automatic speech recognition (ASR), is the process of converting an acoustic signal, captured by a microphone or a telephone, to a text.

The main goal of speech recognition is to get effective ways for mankind to communicate with computers, for example, voice-controlled personal computers. Today's ASR systems are giving considerable compatibility the performance of such systems is far from the perfect system and the research is still gaining on this area.

There are diversified issues concerning the operation of modern ASR systems, such as antiphons which lead to reductions in their efficiency and accuracy to actuate these issues the existence of various forms of variability an articulation in speech [14]. These include Speaker characteristics & variations in acoustic environment,

1.1. Speaker Characteristics

A major factors lead to deformities in the performance of ASR is the articulation environment which leads to commotion such deformity in performances can be boisterous caused by physiological and dialectical differences among the speakers leads to diffident isotopic. Evidence of this can be found by comparing the performance of speaker dependent (SD) and speaker independent (SI) ASR systems. A speaker dependent system is trained using data from a single speaker, and euphony in nature. On the other hand, a speaker independent system is trained using data from a large inconsistent group of different speakers for use by speakers that are not necessarily in the training strata. These techniques can be divided into two main categories: speaker adaptation and speaker normalization.

- Speaker adaptation [11] techniques require the existence of a model or Inunit which has already been trained for single or numerous speakers. The goal of such techniques is to tune the parameters of this model to a new lineament.

  Another technique for adapting HMM-based systems is the maximum likelihood linear regression (MLLR) approach which requires a relatively small amount of adaptation lexicon data. This data is used to compute a number of linear transformations which are applied to the colophon means contained in the model.
- Speaker normalization techniques, [12] on the other hand, instead of replacing the model, perform lingual transformations on the speech signal to compensate for speaker variabilities.

Due to the significant variations in the vocal tract length of different speakers, the positions of the formants produced by different speakers can vary. Therefore, a major category of subtle speaker normalization techniques are focused on normalizing the effective vocal tract length across different vocalist. Vocal tract length normalization (VTLN) and augmented state-space acoustic decoder (MATE) perform this by applying a linear warping to the frequency axis of the utterance, normalizing the position of spectral peaks or formants of speech.

1.2. Variations in Acoustic Environment

Another major factor that leads to degradations in the performance of ASR systems is the presence of noise in the environment. Such degradations in performance can be due to the mismatch between the conditions in which the systems are trained and the ones in which they are operated.

Some speech enhancement approaches are found really good to deal with unknown noise and filtering such as,

- Spectral Subtraction
- Spectral Normalization

The paper is arranged as section II will explain difficulties with ASR, section III gives Classification of Parameters that affect ASR includes Prosody and Environmental parameters, Section IV describes techniques that are helpful to design a robust speech recognition system and finally section V is the conclusion.

2. Difficulties with ASR

The issue of robustness in speech recognition is pregnable of problems. A speech quest may be robust in one environment and yet be impregnable for another and falter less sensitive to noise are not necessarily less sensitive to speaker variability in noise and stress. The main reason for this is that performance of existing recognition systems which may be succinct environment, degrade rapidly in the presence of noise, distortion, and speaker stress.

2.1. Reasons for Difficulties in Speech Recognition:

Difficult problem, largely because of many sources of variability [14, 22] associated with signal

- Acoustic realizations of phonemes, highly dependent on context in which they appear.
  - Phonetic variabilities are exemplified by acoustic differences of phoneme
  - At word boundaries, contextual variations can be quite dramatic
- Acoustic variabilities can result from changes in environment as well as in position and characteristics of transducer.
- The sensitivity to the environment (background noise or channel variability), or
- Within-speaker variabilities can result from changes in person's physical and emotional state, speaking rate, gender, vocal effort, regional accents, speaking style, voice quality etc.
- Differences in sociolinguistic background
- Complexity of the human language, the weak representation of grammatical and semantic knowledge.
- Dialect, and vocal tract size and shape can contribute to across-speaker variabilities.
- Background noise

2.2. Approaches To Improve Speech Recognizers:

There are a variety of approaches that can be used to improve the robustness of speech recognition. These can be classified into five general areas as follows:

- Pre-processing Techniques : Feature Analysis
  - Voice Activity Detector [7]
- Feature Enhancement : front-end signal processing
  - Estimation and noise detection [8]
  - Echo and reverberation detection
  - Normalization
- Feature Extraction : Extracting coefficients
- Model Adaptation: adapting recognition models to the noisy speaker conditions.
- Training Models: consider alternative training using either noisy data, mismatch between training/test data, or modifications which cause the trained models to be more effective for recognizing noisy speech.

3. Classification of Parameters that Affect ASR

Speech Recognition system [16] ebullient in speech processing we regard this as pitch, duration, intensity, voice quality, signal to noise ratio, voice activity detection and strength of Lombard effect these parameters are categorized under two types:

- **Prosody Parameters:** [23] Pitch, duration, intensity and voice quality etc. These parameters are used in speech recognition and especially in the field of speaker characterization. Those systems have to work in general adverse conditions, which leads to the demand of noise robustness for the algorithms estimating the prosodic parameters. Yet it is well known, that most of these parameters are hard to extract from the speech signal, especially under adverse conditions.
- **Environmental Parameters:** the second is the acoustic properties of the environment including the impact on the speaker's voice. As second set we select 'environmental parameters': signal to noise ratio

(SNR), voice activity detection (VAD) [6] and strength of Lombard effect (SLE).

Environmental parameters are used in speech recognition and speaker recognition for noise reduction algorithms. Many approaches exist to estimate SNR and VAD from noisy signals.

The SLE parameters, that decreases the performance of speech recognition systems dramatically with Lombard effect.

3.1. Prosodic Parameters

Speaker variability & characteristics [17] :

the speech signal is non-stationary it not only convey semantic information (the message) but also a lot of information about the speaker himself like, gender, age, social and regional origin, health and emotional state and its identity.

All speakers have their special voices, due to their unique biological body structure (Vocal Track Autonomy) and personality. The voice is not only different among speakers; there are also wide in variations within single specific speaker.

The speaker uniqueness results from a complex combination of physiological and cultural aspects. While finding the variability among speakers through statistical analytic methods found that the first two principal components correspond to the gender and accent respectively. Gender would then appear as the prime factor related to physiological differences, and accent would be one of the most important from the cultural point of view.

The effect of the vocal tract shape on the intrinsic variability of the speech signal between different speakers has play special role in measuring performance of ASR.

Technical Mythology for Compensation Speaker Variation:

Techniques for handling speaker variability are mainly divided into:

- speaker independent feature extraction,
- speaker normalization using The different vocal tract length normalization (VTLN) techniques:
  - speaker-dependent formant mapping
  - transformation of the LPC pole modeling
  - frequency warping, either linear or non-linear
- Speaker adaptation : reduce speaker specificities and tends to further reduce the gap between speaker-dependent and speaker-independent ASR by adapting the acoustic models to a particular speaker

Table 1 : Speaker variations (Please See at the end of the paper)

3.2. Environmental Parameters

The environmental parameters [19] VAD and SNR play a major role in speech recognition system. Extrinsic variabilities are due to the environment: signal to noise ratio may be high but also variable within short time.

Environmental variability & characteristics

- Speech in high noise, with signal-to-noise ratios (SNRs) at or below 0 dB
- Speech in presence of background speech
- Speech in presence of background music
- Speech in highly reverberant environments

Technical Mythology for Compensation of Environmental Variation [21]

General techniques such as

- Compensation,
  - Enhancing speech signal [1],
  - Training models on noisy databases,
  - Designing specific models for noise and speech [3],
  - Considering noise as missing information that can be marginalized in a statistical training of models by making hypotheses on the parametric distributions of noise and speech [9].
- Adaptation [5],
- Multiple models,
- Additional acoustic cues and
- More accurate models

Table 2 : Environmental Variation (Please see at the end of paper)

4. ASR Techniques

In this section, we review methodologies towards improved ASR analysis/modeling accuracy and resistance towards variability sources.

4.1. Front-end Techniques [15]

Feature extraction front-end techniques for the assumption of non-stationary speech signals, high levels of noise, workload task stress, Lombard effect and other variations like,

- Robust features extraction
- The speaker spectral characteristics of speech variability.
- Front-end noise suppression
- Feature compensation techniques for noise reduction.
- Techniques for combining estimation based on different features sets and dimensionality reduction approaches.
- Model adaptation

- Training and Testing in the same conditions.

Feature Extraction Models [13]:

- Mel-Frequency Cepstral Coefficient (MFCC) or Perceptual Linear Prediction (PLP) coefficient, are based on some sort of representation of the smoothed spectral envelope, usually estimated over fixed analysis windows. Such analysis is based on the assumption that the speech signal is quasi-stationary over these segment durations.
- A temporal decomposition technique represents the continuous variation of the LPC parameters as a linearly weighted sum of a number of discrete elementary components. These elementary components are designed such that they have the minimum temporal spread (highly localized in time) resulting in superior coding efficiency.
- A segmental HMM [Achan et al identifies waveform samples at the boundaries between glottal pulse periods with applications in pitch estimation and time-scale modifications.
- The amplitude modulation (AM) and the frequency modulation (FM) used to detect the transition point between the two adjoining QSSs. The power of the residual signal normalized by the number of samples in the window (FM). The AM signal modulates a narrow-band carrier signal (specifically, a monochromatic sinusoidal signal).
- Frequency Scales (M-MFCC, ExpoLog),
- Feature Processing (CMN, VCMN, LP-vs-FFT MFCCs),
- Model Adaptation (PMC), and
- Combinations of gender dependent with gender independent models
- Training and Testing (ANN & HMM).

4.2. Statistical Models:

It is assume that the “clean” speech signal is first passed through a linear filter with unit sample response, whose output is then corrupted by uncorrelated additive noise to produce the degraded speech signal. Under these circumstances, the goal of compensation is, in effect, to undo the estimated parameters characterizing the unknown additive noise and the unknown linear filter, and to apply the appropriate inverse operation.

The popular approaches of spectral subtraction and homomorphic deconvolution are special cases of this model, in which either additive noise or linear filtering effects are considered in isolation. When the compensation parameters are estimated jointly, the problem becomes a nonlinear one, and can be solved using algorithms such as codeword-dependent cepstral normalization (CDCN) and vector-Taylor series compensation (VTS).

4.3. Acoustic Model

The performance of acoustic model is depending on the model matching to the task, which can be obtained through adequate training data and selecting multi-style training.

Model Compensation

- MM decomposition, where dynamic time warping was extended to a 3D-array where the additional dimension represents a noise reference and an optimal path has to be found in this 3D-domain. The major problem was the definition of a local Probability for each box.
- Parallel model decomposition (PMC) where clean speech and noise are both modeled by HMM and where the local probabilities are combined at the level of linear spectrum, this implies that only additive noise can be taken into account.

Adaptation

- A Maximum Likelihood (ML) criterion.

  Try to maximize the probability of a given sequence of observations; Baum-Welch method gives the result.
- A Maximum a Posteriori [2] (MAP)

  In Bayesian [4,10] statistics, a maximum a posteriori probability (MAP) estimate is a mode of the posterior distribution. The MAP can be used to obtain a point estimate of an unobserved quantity on the basis of empirical data. MAP estimation can therefore be seen as a regularization of ML estimation

4.4. Multiple Modeling [10]

Merging too many heterogeneous data in the training corpus makes acoustic models less discriminant. Hence the numerous investigations along multiple modeling, that is the usage of several models for each unit, each model being train from a subset of the training data, defined according to a priori criteria such as gender, age, rate-of-speech (ROS) or through automatic clustering procedures. Ideally subsets should contain homogeneous data, and be large enough for making possible a reliable training of the acoustic models. Gender information is one of the most often used criteria.

- Speaking rate affects notably the recognition performances, thus speaking rate dependent models were studied. It was also noticed that speaking rate dependent models are often getting less speaker-independent because the range of speaking rate shown by different speakers is not the same, and that training procedures robust to sparse data need to be used.
- Signal-to-Noise Ratio (SNR) also impacts recognition performances, hence, besides or in addition to noise reduction techniques, SNR-dependent models have been investigated. In multiple sets of models are trained according to several noise masking levels and the model

set appropriate for the estimated noise level is selected automatically in recognition phase. On the opposite, in acoustic models composed under various SNR conditions are run in parallel during decoding.

- Multi-speaker models [16]: If models of some of the factors affecting speech variation are known, adaptive training schemes can be developed, avoiding training data sparsity issues that could result from cluster-based techniques. This has been used for instance in the case of VTLN normalization, where a specific estimation of the vocal tract length (VTL) is associated to each speaker of the training data. This allows to build a canonical models based on appropriately normalized data. During recognition, a VTL is estimated in order to be able to normalize the feature stream before recognition. More general normalization schemes have also been investigated, based on associating transforms (mostly linear transforms) to each speaker, or more generally, to different cluster of the training data. This transformation can also be constrained to reside in reducing dimensionality eigen space. A technique for factorization selected transformations back in the canonical model is also proposed in, providing a flexible way of building factor specific models, for instance multi-speaker models within a particular noise environment, or multi-environment models for a particular speaker.

4.5. Models for Auxiliary Parameters [20]

Most of speech recognition systems rely on acoustic parameters that represent the speech spectrum, for example cepstral coefficients. However, these features are sensitive to auxiliary information such as pitch, energy, rate-of-speech, etc. the most simple way of using such parameters (pitch and/or voicing) is their direct introduction in the feature vector, along with the cepstral coefficients.

- Pitch has to be taken into account for the recognition of transonic languages. Various coding and normalization schemes of the pitch parameter are generally applied to make it less speaker dependent; the derivative of the pitch is the most useful feature, and pitch tracking and voicing. Pitch, energy and duration have also been used as prosodic parameters in speech recognition systems, or for reducing ambiguity in post-processing steps. Dynamic Bayesian Networks (DBN) offers an integrated formalism for introducing dependence on auxiliary features.
- Speaking rate is another factor that can be taken into account in such a framework. Most experiments deal with limited vocabulary sizes; extension to large vocabulary continuous speech recognition can be achieve through hybrid HMM/BN acoustic modeling.
- TANDEM approach used with pitch, energy or rate of speech. The TANDEM approach transforms the input features into posterior probabilities of sub-word units using artificial neural networks (ANNs), which are then processed to form input features for conventional speech recognition systems.
- Auxiliary Parameters may be used to normalize spectral parameters, for example based on pitch value is used to modify the parameters of the densities (during decoding) through multiple regressions as with pitch and speaking rate.

4.6. Compensation for Environmental Degradation in ASR [18]

Speech samples always affected with the additive noise and linear filtering in normal environment, the use of environmental compensation procedures improves the accuracy in Speech recognition system. The compensation procedures include physiologically-motivated signal processing techniques, modification of either the feature vectors of incoming speech or the internal statistics with which speech recognition systems are trained.

Any change in the environment between the training and testing causes degradation in performance. Continued research is required to improve robustness to new speakers, new dialects, and channel or microphone characteristics, Systems that have some ability to adapt to such changes have to be developed

Some speech enhancement algorithms have proved to be especially important in the development of strategies to cope with unknown noise and filtering.

- Spectral subtraction, to compensate for additive noise. In general, spectral subtraction algorithms attempt to estimate the power spectrum of additive noise in the absence of speech, and then subtract that spectral estimate from the power spectrum of the overall input (which normally includes the sum of speech plus noise). Primarily with the goal of avoiding “musical noise” by “over-subtraction” of the noise spectrum. This method is not appropriate for non-stationary noise, The difficulty to detect pauses (non-speech) in low SNR & musical noise effect. Noise cancellation requires the detection of noise to adaptively extract its spectral and statistical parameters. The ability to discriminate speech from noise enables the calibration of noise cancellation algorithms, identify and filter out these noises from the speech signal. Filtering speech with a high order adaptive FIR filter, when no reference to an external noise source is available, Wiener Filtering for stationary input and noise, no noise reference source is required.
- Spectral Normalization, to compensate for the effects of unknown linear filtering. In general, spectral normalization algorithms first attempt to estimate the average power spectra of speech in the training and testing domains, and then apply the linear filter to the testing speech to “best” convert its spectrum to that of the training speech. Improvements and extensions of

spectral subtraction and spectral normalization algorithms

This section explains some complementary approaches to robust recognition based on initial signal processing techniques.

Approaches to Environmental Compensation

These approaches are grouped as per the effects of noise and filtering.

- Empirical compensation by direct cepstral comparison,
- Model-based compensation, and
- Compensation via cepstral high-pass filtering.

Empirical Compensation by direct cepstral comparison is totally data driven, and requires a "stereo" database that contains time-aligned samples of speech that had been simultaneously recorded in the training environment and in representative testing environments. The success of empirical compensation approaches depends on the extent to which the putative testing environments used to develop the parameters of the compensation algorithm are in fact representative of the actual testing environment.

Empirical Compensation: RATZ and STAR

- The RATZ algorithm modifies the cepstral vectors of incoming speech,
- The STAR algorithm modifies the internal statistical models used by the recognition system.
- RATZ and STAR have a similar conceptual framework.
- RATZ can be considered to be a generalization of algorithms like MFCDCN.
- STAR can be considered to be an extension of the codebook adaptation algorithms.
- RATZ and STAR both assume that the probability density function for clean speech can be characterized as a mixture density. Where the mixture coefficients are fixed for the case of RATZ, and assumed to vary as a function of time to represent the Markov transitional probabilities for the case of STAR.
- Environmental compensation is introduced by modifying the means and variances of the probability density functions.

Model-based compensation assumes a structural model of environmental. Compensation is then provided by applying the appropriate inverse operations. The success of model-based approaches depends on the extent to which the model of degradation used in the compensation process accurately describes the true nature of the degradation to which the speech had been subjected.

Model-Based Compensation: VTS and VPS

- The Vector Taylor Series (VTS) and Vector Polynomial Expansion (VTS) algorithms that develop series approximations to the nonlinear environment function.
- The VTS algorithm approximates the environment function using the first several terms of its Taylor series, where is the vector function evaluated at a particular vector point.
- Similarly, represents the matrix derivative of the vector function at a particular vector point. The higher order terms of the Taylor series involve higher order derivatives resulting in tensors.
- The Taylor expansion is exact everywhere when the order of the Taylor series is infinite.
- VPS approach replaces the Taylor series expansion used in VTS with a more general approach to approximating the environment function.
- VPS is shown to provide a more accurate approximation to the environment function than VTS.
- VPS provided somewhat better recognition accuracy compared to VTS, and at a reduced computational cost.
- It is expected that the difference in error rates between VPS and VTS will increase when implementations of these algorithms that modify the internal statistical models are completed.

Compensation by high-pass filtering implies removal of the steady-state components of the cepstral vector. The amount of compensation provided by high-pass filtering is more limited than the compensation provided by the two other types of approaches, but the procedures employed are simple and effective that they should be included in virtually every current speech recognition system.

Cepstral High-Pass Filtering: RASTA and CMN

- In Relative Spectral Processing or RASTA processing, a high-pass (or band-pass) filter is applied to a log-spectral representation or cepstral representation of speech.
- Cepstral mean normalization (CMN) is an alternate way to high-pass filter cepstral coefficients.
- High-pass filtering in CMN is accomplished by subtracting the short-term average of cepstral vectors from the incoming cepstral coefficients.
- RASTA and CMN are effective in compensating for the effects of unknown linear filtering in the absence of additive noise because under these circumstances the ideal cepstral compensation vector is a constant that is independent of SNR and VQ cluster identity. Such a compensation vector is, in fact, equal to the long-term average difference between all cepstra of speech in the training and testing environments.
- The high-pass nature of both the RASTA and CMN filters forces the average values of cepstral coefficients to be zero in the training and testing environments

individually, which, i, implies that the average cepstra in the two environments are equal to each other.

- Cepstral high-pass filtering can also be thought of as a degenerate case of compensation based on direct cepstral comparison.

Joint compensation for the effects of noise and filtering has proceeded in two phases.

- In the initial phase concerned with understanding the basic properties of the environment function and with the development of compensation procedures that were relatively simple but that provided significant improvements in recognition accuracy compared to the accuracy that could be obtained from independent compensation for the effects of noise and filtering.
- During the second phase of algorithm development focused on the development of algorithms that could achieve greater recognition accuracy under the most arduous conditions through the use of more accurate mathematical characterizations of the effects of noise and filtering.

Table 1 : Speaker Variations

Reason for Variation	Effect of Variation	A General Technique for handling Speech variation
Anatomy of vocal tract	The power spectral density of speech varies over time according to the glottal signal and the configuration of the speech articulators.	▪ Compensation and invariance (Normalization) ▪ Vocal Tract Length Normalization (VTLN) ▪ Hidden Markov Models (HMMs), as a sequence of stationary random regimes
Realization	Speaker can not produce the same acoustic wave for the same word if he/she pronounced over and over again	▪ Auto corrélation technique ▪ Clustering techniques ▪ Speeker Adaption model
Ambiguity : Homophones Ambiguity : Word boundary	• Words that sound the same, but have different orthography • Multiple ways of grouping phones into words.	▪ Language Model ▪ Use of multiple acoustic models associated to large groups of pronunciation variants (lexical level) speakers
The sex of the speaker	In general Women have shorter vocal tract than men. The fundamental tone of women’s voices is roughly two times higher than men’s	▪ Vocal Tract Length Normalization (VTLN) ▪ Cepstral Mean subtraction (CMS) ▪ Mean & Variance Normalization (MVN) ▪ Spectral Normalization
Speaking rate & Speaking style	The spectral effects of speech rate variations.	▪ Speech rate estimator ▪ The evaluation of the frequency of phonemes or syllables in a sentence ▪ Normalization by dividing the measured phone duration by the average duration of the underlying phone
Regional and Social Dialects	Dialects are group related variation within a language. Regional dialect involves features of pronunciation, vocabulary and grammar which differ according to the geographical area. Social dialects are distinguished by features of pronunciation, vocabulary and grammar according to the social group of the speaker.	▪ Consider dialects as another language in ASR, due to the large differences between two dialects.
Amount of data and search space	The quality of the speech signal decreases with a lower sampling rate, resulting in incorrect analysis. Minimizing lexicon (set of words) causes out-of-vocabulary	▪ Use of large vocabulary
Foreign and	Variations in speaker accent degrade the	▪ Select an appropriate language model or adapt to

Regional Accents	performance of speech recognition systems that fails to recognize target language.	the accent/speaker ▪ Recognizer should train on target language. ▪ Use of multiple acoustic models associated to large groups of pronunciation variants (lexical level) speakers. ▪ Adapt of Multilingual phone models
Age	• The difference in vocal tract size results in a non-linear increase of the formant frequencies. • Larger spectral and supra-segmental variations and wider variability in formant locations and fundamental frequencies in the speech signal.	▪ Use larges size of the pronunciation dictionary, corpora for children and adults ▪ Selection of language models which are customized for children speech & physiological besides. ▪ Adapting the acoustic features of children speech to match that of acoustic models trained from adult speech. ▪ Use of vocal tract length normalization (VTLN) and spectral normalization approaches
Emotions	Emotions in speech recognition is concentrated on attempting to classify a “stressed” speech signal into its correct emotion category. Intrinsic variabilities: loud, soft, Lombard, fast, angry, scared; and noise.	▪ Improved front-end processing, feature extraction methods for the recognition of stressed and non-stressed speech simultaneously. ▪ Improved back-end processing or robust recognition measures. ▪ Improved training methods: Multi-style training and simulated stress token generation.
Dis-fluencies in speech	False starts, Repetitions, Hesitations and filled pauses, Slips of the tongue etc.	▪ Spectral Subtraction methods ▪ Improved feature extraction methods ▪ Appropriate training model

Table 2 : Environmental Variation

Reason for Variation	Effect of Variation	Technique for handling Speech variation
Noise	Unwanted information in the speech signal like voices in the background that corrupts the quality of speech signal and degrades the performance of ASR system.	▪ Spectral Subtraction Method ▪ Noise Estimation, Cancellation approaches and filters ▪ A high order adaptive FIR filter. When no reference to an external noise source is available, ▪ Wiener Filtering for stationary input and noise, no noise reference source is required. ▪ SPLICE algorithm works on spectral representation ▪ ALGONQUIN algorithm works on log-spectra
Echo effect	The speech signal bounced on some surrounding object, and that arrives in the microphone a few milliseconds later. This echo effect adds with original speech signal, and difficult to get clean original speech.	▪ Echo Cancellation Algorithms ▪ Least-Mean-Square (LMS) and ▪ Normalized LMS (NLMS) ▪ Approach for Echo Cancellation ▪ Minimum Statistics (MS) ▪ Eigenvalue Decomposition ▪ Fourier Transform (DFT) ▪ State-Space Model ▪ Vector Taylor Series (VTS) ▪ ALGONQUIN Method ▪ Time-Variant Estimate

		▪ Switching Linear Dynamic Model (SLDM) ▪ Bayesian Estimation Framework ▪ Random-Walk State Model
Reverberation	If the place in which the speech signal has been produced is strongly echoing, then this may give raise to a phenomenon called reverberation, which may last even as long as seconds. The original speech signals mask with echoing.	▪ Additive noise filtering algorithms ▪ Adaptive Schemes ▪ Proportionate Schemes ▪ Proportionate Adaptive Filters ▪ Block-Based Combination ▪ Combination Schemes ▪ Subband Adaptive Filtering ▪ Uniform Over-Sampled DFT Filter Banks ▪ Subband Over-Sampled DFT filter banks (FB) ▪ Time-Domain Considerations ▪ Volterra filters ▪ Proportionate-Type Algorithms ▪ Sparseness-Controlled Algorithms
Channel Variability	The noise that changes over time, and different kinds of microphones and everything else that affects the content of the acoustic wave from the speaker to the discrete representation in a computer.	▪ Cepstral Mean Subtraction ▪ The RASTA filtering of spectral trajectories.
Convolution Noise	Speech signal quality degradations due to the channel come from its slowly varying spectral properties (or impulse response).	▪ Averaging speech features (Cepstral Mean Subtraction) ▪ Evaluating the impulse response as missing data and combined with additive noise reduction. ▪ Low pass filtering, by removing Cepstral mean from all feature vectors of the utterance.
Lombard effect	Due to noisy environments acoustic correlates in the speech signal. But to quantify this effect no specification is known.	▪ Additive noise filtering algorithms ▪ Applying Low pass and High pass filters
Physical Stress The force environment, Auditory distraction, Thermal environment, Personal equipment. Emotional Stress Task load, Mental fatigue, Mission anxieties and Background anxieties.	The noise can be considered stationary during a vocal command, but from one vocal command to another, its characteristics can change. .	▪ Applying suitable feature extraction method like LPCC, MFCC ▪ Noise estimation and cancellation algorithms. ▪ Noise cancellation to be performed by Wiener type Filtering

5. Discussion & Conclusion

ASR is a challenging task. In this paper, we have addressed some of the difficulties of speech recognition, the most problematic issues being the large search space and the strong variability, this covers accent, speaking rate & style, regional and social dialects, speaker physiology, age, emotions etc.

This paper covered the different causes of acoustical and environmental variability. There are some attributes of the environment that remain relatively constant through the course of an utterance such as the recording equipment, the amount of room reverberation, and the acoustical characteristics of the particular speaker using the system. Other factors, like the noise and signal levels, will be assumed to vary slowly compared to the rate at which speech changes.

Conventional techniques that compensate for the effects of additive noise and linear filtering of speech sounds can provide substantial improvement in recognition accuracy when the cause of the acoustical degradation is quasi-stationary. The recognition of speech at lower SNRs, and especially speech in the presence of transient sources of interference including especially background speech and background music remain essentially unsolved problems at present.

Some techniques are explained for Environment compensation, which remove speech variabilities due to environment and channel characteristics, speaker normalization techniques, which remove variabilities due to speaker characteristics, and discriminant feature space-transformation techniques, which are aimed at increasing the class discrimination of the speech data.

Finally, the paper proposed an overview of general techniques for better handling intrinsic and extrinsic variation sources in ASR, mostly tackling the speech analysis and acoustic modeling aspect.

6. References

[1] Philipos C, Reasons Why Current Speech-Enhancement Algorithms Do Not Improve Speech Intelligibility And Suggested Solutions, IEEE Transactions On Audio, Speech, And Language Processing, Vol. 19, No. 1, January 2011

[2] Suhadi Suhadi, A Data-Driven Approach To A Priori Snr Estimation, IEEE Transactions On Audio, Speech, And Language Processing, Vol. 19, No. 1, January 2011,

[3] Ke Hu, Unvoiced Speech Segregation From Nonspeech Interference via CASA and Spectral Subtraction, IEEE Transactions On Audio, Speech, And Language Processing, Vol. 19, No. 6, August 2011

[4] Antonio Miguel, Bayesian Networks for Discrete Observation Distributions in Speech Recognition, IEEE Transactions On Audio, Speech, And Language Processing, Vol. 19, No. 6, August 2011

[5] Sayed. A. A Family Of Adaptive Filter Algorithms In Noise Cancellation For Speech Enhancement, International Journal Of Computer And Electrical Engineering, Vol. 2, No. 2, April 2010. 1793-8163

[6] Urmila Shrawankar, Dr. V M Thakare, Voice Activity Detector and Noise Trackers for Speech Recognition System in Noisy Environment, International Journal of Advancements in Computing Technology (IJACT), 2010, ISSN: 2005-8039

[7] C. Ganesh Babu, Performance analysis of voice activity detection algorithm for robust speech recognition system under different noisy environment, Journal of scientific and Industrial research Vol 69, July 2010, PP 515-522

[8] Urmila Shrawankar, Dr. V M Thakare, “Noise Estimation and Noise Removal Techniques for Speech Recognition in Adverse Environment”, Springer-IIP2010, Manchester , UK , October 13-16, 2010

[9] Bj¨orn Schuller, Recognition of Noisy Speech: A Comparative Survey of Robust Model Architecture and Feature Enhancement, EURASIP Journal on Audio, Speech, and Music Processing, Volume 2009, Article ID 942617

[10] Jiucang Hao, Speech Enhancement, Gain, And Noise Spectrum Adaptation Using Approximate Bayesian Estimation, IEEE Transactions On Audio, Speech, And Language Processing, Vol. 17, No. 1, January 2009

[11] Mohamed Benzeghiba, Impact of variabilities on speech recognition, SPECOM'2006, 11th International Conference Speech and Computer, June 25-29, 2006, Saint-Petersburg, Russia

[12] STERN, R. M. “Signal Separation Motivated by Human Auditory Perception: Applications to Automatic Speech Recognition,.” Chapter in Speech Separation by Humans and Machines, P. Divenyi, Ed., Springer-Verlag, 2004

[13] Ivanov, Alexei V. / Petrovsky, Alexander A. (2004): "Anthropomorphic feature extraction algorithm for speech recognition in adverse environments", In SPECOM-2004, 166-173

[14] Markus Forsberg, Why Is Speech Recognition Difficult? Chalmers University of Technology, 2003 – Citeseer, February 24, 2003

[15] Hansen, John H. L. / Sarikaya, Ruhi / Yapanel, Umit / Pellom, Bryan (2001): "Robust speech recognition in noise: an evaluation using the SPINE corpus", In EUROSPEECH-2001, 905-911.

[16] Timothy R. Anderson, Applications Of Speech-Based Control, RTO Lecture Series, held in Bre’tigny, France, 7-8 October 1998, and in Ohio, USA, 14-15 October 1998, and published in RTO EN-3.

[17] Timothy R. Anderson, APPLICATIONS OF SPEECH-BASED CONTROL, the RTO Lecture Seeris on “Alternative Control Technologies: Human Factors Issues”, Bre’tigny, France, 7-8 October 1998, and in Ohio, USA, 14-15 October 1998, and published in RTO EN-3.

[18] STERN, R. M., RAJ, B., and MORENO, P. J., “Compensation for Environmental Degradation in Automatic Speech Recognition,.”Channels, April, 1997, Pont-au-Mousson, France, pp. 33-42.

[19] Carlos Avendano and Hynek Hermansky, "On the Properties of Temporal Processing for Speech in Adverse Environments," WASPA’97, Mohonk, NY 1997.

[20] Adam L. Buchsbaum, Raffaele Giancarlo, Algorithmic Aspects in Speech Recognition: An Introduction, Journal of Experimental Algorithmics (JEA), Volume 2, 1997

[21] J.-P. HATON, Problems And Solutions For Noisy Speech Recognition, Journal De Physique IV, 1994

[22] Larry E. Humes, Lisa Roberts, Speech-Recognition Difficulties Of The Hearingimpaired Elderly: The Contributions Of Audibility, Journal Of Speech And Hearing Research, Volume 33, 726-735, December 1990

[23] Harald Höge, Basic Parameters In Speech Processing, The Need For Evaluation

Query Optimization Using Genetic Algorithms in the Vector Space Model

Eman Al Mashagba[1], Feras Al Mashagba[2] and Mohammad Othman Nassar[3]

[1] **Computer Information Systems, Irbid Private University, Irbid, 22110, Jordan**

[2] **Computer Information Systems, Amman Arab University, Amman, 11942, Jordan**

[3] **Computer Information Systems, Amman Arab University, Amman, 11942, Jordan**

Abstract

In information retrieval research; Genetic Algorithms (GA) can be used to find global solutions in many difficult problems. This study used different similarity measures (Dice, Inner Product) in the VSM, for each similarity measure we compared ten different GA approaches based on different fitness functions, different mutations and different crossover strategies to find the best strategy and fitness function that can be used when the data collection is the Arabic language. Our results shows that the GA approach which uses one-point crossover operator, point mutation and Inner Product similarity as a fitness function is the best IR system in VSM.

Keywords: *information retrieval, vector space model, query optimization, genetic algorithms.*

1. Introduction

Information retrieval (IR) can be defined as the study of how to determine and retrieve from a corpus of stored information the portions which are responsive to particular information needs [1]. IR is also concerned with text representation, text storage, text organization, and the retrieval of stored information items that are similar in some sense to information requests received from users.The major information retrieval model includes: the vector space model, Boolean model, Fuzzy sets model and the probabilistic retrieval model. These models are used to find the similarity between the query and the documents in order to retrieve the documents that reflect the query. Vector space model usually use Cosine, DICE, Jaccard, or Inner Product as a similarity measures. The similarity then used to evaluate the effectiveness of IR system using two measures: Precision which is a ratio that compares the number of relevant documents found to the total number of returned documents [2], and Recall which is the system's ability to retrieve all related documents of a query [2]. The problem with the IR models is that it may converge to a result that is only locally optimal, which means it may lead to form a query that is better than the original form but significantly poorer than another undetected form, so GA can be used to solve this problem.

A GA is an adaptive heuristic search algorithm premised on the evolutionary ideas of natural selection and genetics [3]. The GA approach has gained importance and popularity, as evident in the number of studies that have used it to improve different optimization procedures to be able to find a global solution in many problems. GA have been used for difficult problems (such as NP-hard problems), for machine learning and also for evolving simple programs. In this paper; and for each similarity measure (Dice, and Inner Product) in the vector space model we will implement and compare ten different genetic algorithms settings (different mutation techniques, different fitness functions, different crossover techniques) to optimize the user query. As a test bed; we are going to use an Arabic data collection which was presented for the first time by [16]; this data set is composed from 242 documents and 59 queries, the correct answer for each query (relevant documents) is also known in advanced.

Arabic is the official language of over than twenty one Arab countries, and it is the religious language of more than one billion Muslims around the world. The Arabic language is unique and difficult language; the difficulty comes from several sources; amongst them: it differs syntactically, morphologically, and semantically from other Indo-European languages [13]. Compared to English, Arabic language is more sparsed, which means that for the same text length English words are repeated more often than Arabic words [14, 15]. Sparseness may negatively affect the retrieval quality in Arabic language because Arabic terms will get less weight compared to English. In written Arabic, most letters take many forms of writing. Also, there is a punctuation associated with some letters that may change the meaning of two identical words. Finally; comparing to English roots, Arabic roots are more complex. The same Arabic root, depending on the context, may be derived from multiple Arabic words.

The uniqueness and the special properties for the Arabic language, its differences from the English and the other languages, and the lack of similar studies in the literature was our motivator to conduct a deep and rich comparative study based on Arabic data collection.

.

2. Previous Studies

There are several studies that used GA in information retrieval systems to optimize the user query based on English data collections such as [4, 5, 7, 6, 7, 5, 9, 10, 11, 12, 18].

In their experiments for the VSM [8,4,6], the authors presents many methods: the connectionist Hopfield network; the symbolic ID3/ID5R, evolution- based genetic algorithms, symbolic ID3 Algorithm, evolution-based genetic algorithms, Simulated Annealing, neural networks, genetic programming. They found that these techniques are promising in their ability to analyze user queries, identify users' information needs, and suggest alternatives for search. In [9, 11, 7, 5,12] the VSM have been used, different mutation probabilities, new crossover operation, new fitness functions for the GA have been tested to improve the IR performance. Mercy and Naomie [10] propose a framework of data fusion approach based on linear combinations of retrieval status values obtained from Vector Space Model and Probability Model system. They used Genetic Algorithm (GA) to find the best linear combination of weights assigned to the scores of different retrieval system to get the most optimal retrieval performance.

Using GA to improve the performance of Arabic information system is rare in the literature. In [17] the researchers used Genetic Algorithms to improve performance of Arabic information retrieval system, which based on vector space model. The performance was enhanced through the usage of an adaptive matching function, which obtained from a weighted combination of four similarity measures (Dot, Cosine, Jaccard and Dice).

As we can see from the previous studies; a little research have Been conducted for the Arabic data collections, and since the information retrieval (IR) is one of the most crucial components in search engines and their optimization would have a great effect on improving the searching efficiency, it is important to conduct a comprehensive comparison between different genetic algorithms settings for each similarity measure in the VSM to decide which GA setting and which similarity measure is more useful when used with the Arabic data collections.

3. Vector Space Model (VSM)

The vector space model (VSM) is an IR model that represents the documents and queries as vectors in a multidimensional space, whose dimensions are the terms used to build an index to represent the documents. Lexical scanning is required in order to identify the terms, then an optional stemming process applied to the words and then the occurrence of those stems is computed. Finally the query and the document vectors are compared using different similarity measures (e.g. Cosine, DICE, Jaccard, and Inner Product), those similarity measures are shown in Table 1. In vector space model the importance of terms is determined by their weights, which are computed by using the statistical distributions of the terms in the collection and in the documents. Where $W_{i,j}$ in table 1 are the weights of the ith term in document j, and in the query respectively.

4. Genetic Algorithms (GA)

The basic concept of GA is designed to simulate processes in natural systems necessary for evolution. As such they represent an intelligent exploitation of a random search within a defined search space to solve a problem.

GAs exploits the idea of the survival of the fittest and an interbreeding population to create a novel and innovative search strategy. A population of strings, representing solutions to a specified problem, is maintained by the GA. The GA then iteratively creates new populations from the old by ranking the strings and interbreeding the fittest to create new strings, which are hopefully closer to the optimum solution to the problem at hand. So in each generation, the GA creates a set of strings from the bits and pieces of the previous strings. The idea of survival of the fittest is of great importance to genetic algorithms. GAs use what is termed as a fitness function in order to select the fittest string that will be used to create new, and conceivably better, populations of strings. The only thing that the fitness function must do is to rank the strings in some way by producing the fitness value. These values are then used to select the fittest strings.

Table 1: Different Similarity Measures.

Similarity Measure	Evaluation for Binary Term Vector	Evaluation for Weighted Term Vector
Cosine	$sim(d,q) = 2\frac{\lvert d \cap q \rvert}{\lvert d \rvert^{1/2} \bullet \lvert q \rvert^{1/2}}$	$sim(d_j, q) = \frac{\sum_{i-1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i.j}^2} \times \sqrt{\sum_{j=1}^{t} w_{i,q}^2}}$
Dice	$sim(d,q) = 2\frac{\lvert d \cap q \rvert}{\lvert d \rvert + \lvert q \rvert}$	$sim(d_j, q) = \frac{2\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sum_{i=1}^{t} w_{i,j}^2 + \sum_{i=1}^{t} w_{i,q}^2}$
Jaccard	$sim(d,q) = \frac{\lvert d \cap q \rvert}{\lvert d \rvert + \lvert q \rvert - \lvert d \cap q \rvert}$	$sim(d_j, q) = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sum_{i=1}^{t} w_{i,j}^2 + \sum_{i=1}^{t} w_{i,q}^2 - \sum_{i=1}^{t} w_{i,j} \times w_{i,q}}$
Inner Product	$\lvert d_i \cap q_k \rvert$	Sim= $\sum_{k=1}^{t} (d_{ik} \bullet q_k)$

The GA algorithm flowchart is illustrated in Figure 1. Genetic algorithm operations can be used to generate new and better generations. As shown in Figure 1 the genetic algorithm operations include:

A. Reproduction: the selection of the fittest individuals based on the fitness function.

B. Crossover: the exchange of genes between two individual chromosomes that are reproducing. In one point cross over a chunk of connected genes will be swapped between two chromosomes. There are many crossover strategies such as n-point crossover [11], restricted crossover [7], uniform crossover [30], fusion operator [7] and dissociated crossover [7]. For more details about the crossover strategies you can see the related references.

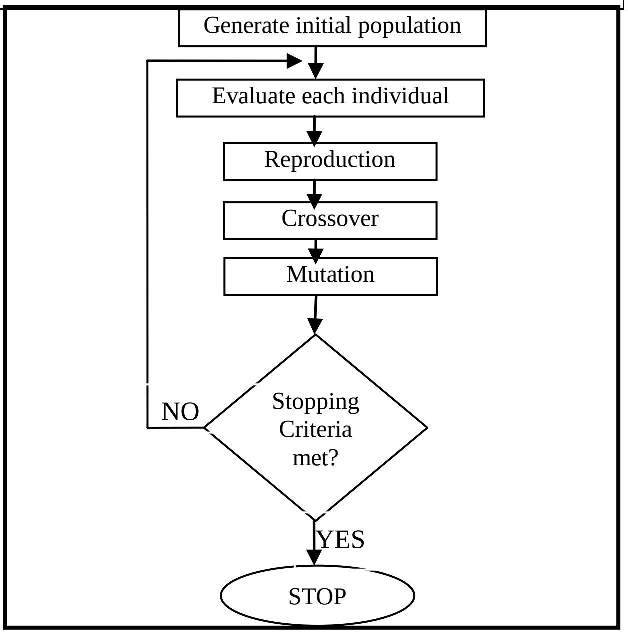


Fig. 1: Flowchart for Typical Genetic Algorithm.

C. Mutation: is the process of randomly altering the genes in a particular chromosome. There are two types of mutation:

1) Point mutation: in which a single gene is changed.
2) Chromosomal mutation: where some number of genes is changed completely.

5. Experiment

In this study we used IR system based on VSM model that was built and implemented by Hanandeh [6] to handle the 242 Arabic abstracts collected from the Proceedings of the Saudi Arabian National Conference [16]. In this study the significant terms are extracted from relevant and irrelevant documents then assigned weights. The binary weights of the terms are formed as a query vector, and then the query vector is adapted as a chromosome. Finally the GA is applied to get an optimal or near optimal query vector, and the result of the GA approach is compared with the result of the traditional IR system without using a GA.

This study was conducted as following:

1) Representation of the chromosomes: The chromosomes are represented as following:

   a) Binary representation: The chromosomes use a binary representation, and are converted to a real representation by using a random function.
   b) Number of Genes: We will have the same number of genes as the query and the feedback documents that have terms with non-zero weights.
   c) Chromosome size: The size of the chromosomes will be equal to the number of terms of the set (feedback documents+ the query set).
   d) The query vector: The query is represented as a binary.
   e) Terms update: Terms are modified by applying the random function on the terms weights.
   f) GA approach: The GA approaches receive an initial population chromosomes corresponding to the top 15 documents retrieved from traditional IR with respect to that query.

2) Fitness Function: Fitness function is a performance measure or reward function which evaluates how each solution is good. In this study Dice, and Inner Product similarity measures are used as fitness functions

3) Selection: Chromosomes selection depends on the fitness function where the higher values have a higher probability to be selected in the next generation.

4) Operators: In our GA approaches, we use two GA operators to produce offspring chromosomes, which are:

1. Crossover: it is the genetic operator that mixes two chromosomes together to form new offspring. In this experiment crossover occurs only with crossover probability Pc (Pc=0.8). Chromosomes are not subjected to crossover remain unmodified. Higher fitness chromosome has an opportunity to be selected more than lower ones, so good solution always survives to the next generation. In this study; different crossover strategies were used for VSM :
   a) One-point crossover operator.
   b) Restricted crossover operator.
   c) Uniform crossover operator.
   d) Fusion operator.
   e) Dissociated crossover.

2. Mutation which involves the modification of the gene values of a solution with some probability Pm. chromosome modification using mutation may lead to better or poorer chromosomes. If they are poorer than old chromosome they are eliminated in selection step. In this experiment we used a mutation probability (Pm=0.7) and two different mutation strategies:
   a) Point mutation.
   b) Chromosomal mutation.

Putting the previous details together we created a number of GA strategies. Those strategies will be used with each similarity measure (Dice, and Inner Product) in the VSM, Those strategies are as following:

1) GA1: GA that use one-point crossover operator and point mutation.
2) GA2: GA that use one-point crossover operator and chromosomal mutation.
3) GA3: GA that use restricted crossover operator and point mutation.
4) GA4: GA that use restricted crossover operator and chromosomal mutation.
5) GA5: GA that use uniform crossover operator and point mutation.
6) GA6: GA that use uniform crossover operator and chromosomal mutation.

7) GA7: GA that use fusion operator and point mutation.
8) GA8: GA that use fusion operator and chromosomal mutation.
9) GA9: GA that use dissociated crossover and point mutation.
10) GA10: GA that use dissociated crossover and chromosomal mutation.

6. Results for the GA strategies Using Dice Similarity

The results for the GA strategies using Dice similarity are shown in Table 2 and Table 3. From those tables we notice that GA1, GA2, GA4, GA5, GA7, GA8, GA9 and GA10 give a high improvement than traditional IR system with 2.726679%, 4.256249%, 3.051032%, 5.940507%, 5.98964%, 6.095792%, 10.83388% and 9.757293% respectively while GA3 and GA6 give a low improvement than traditional IR system with -1.19504% and -4.68231% respectively. Which means that GA9 that use dissociated crossover and point mutation gives the highest improvement over the traditional approach with 10.83388%.

7. Results for the GA strategies Using Inner Product Similarity

The results for the GA strategies using Inner product similarity are shown in Table 4 and Table 5. From those tables we notice that GA1, GA2, GA3, GA4, GA5,GA9 and GA10 give a high improvement than traditional IR system with 11.9444%, 3.355853%, 3.271745%, 3.203264%, 2.912908%, 5.074422% and 6.307254% respectively while GA6, GA8 and GA9 give a low improvement than traditional IR system with -2.71346%, -2.32334% and -3.60072% respectively. This means that GA1 that use GA that use one-point crossover operator and point mutation gives the highest improvement over the traditional approach with 11.9444%.

8. Comparison between the Best GA's Strategies

Table 6 shows the comparison between Dice (GA9) and Inner Product (GA1). From this table we notice that the Inner Product (GA1) is better than Dice (GA9) in all recall levels. Which means that Inner Product(GA1) that use one-point crossover operator and point mutation and use Inner Product similarity as a fitness function represent the best IR system in VSM to be used with the Arabic data collection.

Table 6: Comparison Between the Best GA Strategies (Each Similarity Measures).

Recall	Dice(GA9)	Inner Product(GA1)
0.1	0.141	0.146
0.2	0.197	0.208
0,3	0.298	0.301
0.4	0.277	0.283
0.5	0.402	0.405
0.6	0.408	0.409
0.7	0.396	0.413
0.8	0.412	0.437
0.9	0.441	0.487
Average	0.330222	0.343222

9. Conclusions

For each similarity measure (DICE and Inner Product) in the VSM we compared ten different GA approaches, and by calculating the improvement of each approach over the traditional IR system, we noticed that most approaches (GA1, GA2, GA4, GA5, GA8, GA9 and GA10) give improvements compared to the traditional IR system, also we noticed that in the inner product the one-point crossover operator and point mutation gives the highest improvement over the traditional approach.

Table 2: Average Recall and Precision Values for 59 Query by Applying GA's on Dice Similarity.

Recall	Dice	GA1	GA2	GA3	GA4	GA5	GA6	GA7	GA8	GA9	GA10
0.1	0.131	0.133	0.143	0.129	0.132	0.134	0.136	0.146	0.134	0.141	0.139
0.2	0.172	0.173	0.177	0.165	0.183	0.191	0.187	0.187	0.182	0.197	0.193
0,3	0.262	0.266	0.268	0.254	0.277	0.288	0.177	0.274	0.285	0.298	0.297
0.4	0.214	0.233	0.225	0.212	0.221	0.242	0.232	0.229	0.254	0.277	0.278
0.5	0.357	0.367	0.387	0.363	0.366	0.376	0.223	0.393	0.399	0.402	0.393
0.6	0.379	0.388	0.389	0.385	0.389	0.384	0.391	0.387	0.389	0.408	0.397
0.7	0.383	0.389	0.401	0.375	0.386	0.387	0.386	0.399	0.387	0.396	0.401
0.8	0.388	0.395	0.399	0.387	0.398	0.403	0.394	0.405	0.402	0.412	0.414
0.9	0.431	0.446	0.432	0.422	0.443	0.455	0.437	0.437	0.432	0.441	0.431
Average	**0.3018**	**0.31**	**0.3134**	**0.2991**	**0.3105**	**0.3177**	**0.2847**	**0.3174**	**0.3182**	**0.3302**	**0.327**

Table 3: GA's Improvement in Dice Similarity (GA's Improvement %).

Recall	GA1	GA2	GA3	GA4	GA5	GA6	GA7	GA8	GA9	GA10
0.1	1.526718	9.160305	-1.52672	0.763359	2.290076	3.816794	11.45038	2.290076	7.633588	6.10687
0.2	0.581395	2.906977	-4.06977	6.395349	11.04651	8.72093	8.72093	5.813953	14.53488	12.2093
0,3	1.526718	2.290076	-3.05344	5.725191	9.923664	-32.4427	4.580153	8.778626	13.74046	13.35878
0.4	8.878505	5.140187	-0.93458	3.271028	13.08411	8.411215	7.009346	18.69159	29.43925	29.90654
0.5	2.80112	8.403361	1.680672	2.521008	5.322129	-37.535	10.08403	11.76471	12.60504	10.08403
0.6	2.37467	2.638522	1.583113	2.638522	1.319261	3.166227	2.110818	2.638522	7.651715	4.74934
0.7	1.56658	4.699739	-2.08877	0.78329	1.044386	0.78329	4.177546	1.044386	3.394256	4.699739
0.8	1.804124	2.835052	-0.25773	2.57732	3.865979	1.546392	4.381443	3.608247	6.185567	6.701031
0.9	3.480278	0.232019	-2.08817	2.784223	5.568445	1.392111	1.392111	0.232019	2.320186	0
Average	**2.726679**	**4.256249**	**-1.19504**	**3.051032**	**5.940507**	**-4.68231**	**5.98964**	**6.095792**	**10.83388**	**9.757293**

Table 4: Average Recall and Precision Values for 59 Query by Applying GA's on Inner Product Similarity.

Recall	Dice	GA1	GA2	GA3	GA4	GA5	GA6	GA7	GA8	GA9	GA10
0.1	0.132	0.146	0.134	0.146	0.134	0.135	0.139	0.134	0.129	0.139	0.135
0.2	0.178	0.208	0.182	0.187	0.191	0.185	0.186	0.167	0.169	0.192	0.192
0,3	0.265	0.301	0.285	0.274	0.288	0.288	0.177	0.256	0.272	0.268	0.287
0.4	0.221	0.283	0.254	0.229	0.242	0.242	0.227	0.223	0.211	0.231	0.255
0.5	0.376	0.405	0.399	0.393	0.376	0.376	0.366	0.365	0.344	0.399	0.399
0.6	0.381	0.409	0.389	0.387	0.384	0.384	0.391	0.377	0.371	0.408	0.397
0.7	0.391	0.413	0.387	0.399	0.387	0.387	0.386	0.389	0.386	0.411	0.404
0.8	0.394	0.437	0.402	0.405	0.403	0.403	0.394	0.386	0.393	0.425	0.422
0.9	0.456	0.487	0.432	0.437	0.455	0.455	0.445	0.423	0.408	0.459	0.466
Average	**0.3104**	**0.3432**	**0.3182**	**0.3174**	**0.3177**	**0.3172**	**0.3012**	**0.3022**	**0.2981**	**0.3257**	**0.328556**

Table 5: GA's Improvement in Inner Product Similarity (GA's Improvement %).

Recall	GA1	GA2	GA3	GA4	GA5	GA6	GA7	GA8	GA9	GA10
0.1	10.60606	1.515152	10.60606	1.515152	2.272727	5.30303	1.515152	-2.27273	5.30303	2.272727
0.2	16.85393	2.247191	5.05618	7.303371	3.932584	4.494382	-6.17978	-5.05618	7.865169	7.865169
0,3	13.58491	7.54717	3.396226	8.679245	8.679245	-33.2075	-3.39623	2.641509	1.132075	8.301887
0.4	28.0543	14.93213	3.61991	9.502262	9.502262	2.714932	0.904977	-4.52489	4.524887	15.38462
0.5	7.712766	6.117021	4.521277	0	0	-2.65957	-2.92553	-8.51064	6.117021	6.117021
0.6	7.349081	2.099738	1.574803	0.787402	0.787402	2.624672	-1.04987	-2.62467	7.086614	4.199475
0.7	5.626598	-1.02302	2.046036	-1.02302	-1.02302	-1.27877	-0.51151	-1.27877	5.11509	3.324808
0.8	10.91371	2.030457	2.791878	2.284264	2.284264	0	-2.03046	-0.25381	7.86802	7.106599
0.9	6.798246	-5.26316	-4.16667	-0.2193	-0.2193	-2.41228	-7.23684	-10.5263	0.657895	2.192982
Average	**11.9444**	**3.355853**	**3.271745**	**3.203264**	**2.912908**	**-2.71346**	**-2.32334**	**-3.60072**	**5.074422**	**6.307254**

References

[1] Tengku M.T., Sembok, C.J., and van Rijsbergen, "A simple logical-linguistic document retrieval system", Information Processing & Management, Volume 26, Issue 1, pp. 111-134, 1990.

[2] J. Carlberger, H. Dalianis, M. Hassel, O. Knutsson, "Improving Precision in Information Retrieval for Swedish using Stemming", In the Proceedings of NoDaLiDa-01 - 13th Nordic Conference on Computational Linguistics, May 21-22, 2001, Uppsala, Sweden.

[3] Goldberg, D. E., Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, 1989.

[4] Hsinchun C., "Machine Learning for Information Retrieval: Neural Networks, Symbolic Learning, and Genetic Algorithms", Journal of the American Society for Information Science. Volume 46 Issue 3, April 1995.

[5] D. Vrajitoru, “Crossover improvement for the genetic algorithm in information retrieval”, Information Processing& Management, 34(4), pp. 405–415, 1998.

[6] Hsinchun C, Ganesan S, Linlin S, "A Machine Learning Approach to Inductive Query by Examples: An Experiment Using Relevance Feedback, ID3, Genetic Algorithms, and Simulated Annealing", Journal Of The American Society For Information Science. 49(8):693–705, 1998.

[7] Vicente P., Cristina P., "Order-Based Fitness Functions for Genetic Algorithms Applied to Relevance Feedback", Journal Of The American Society For Information Science And Technology, 54(2):152–160, 2003.

[8] Andrew T., "an Artificial Intelligence Approach to Information Retrieval", Information Processing and Management, 40(4):619-632, 2004.

[9] Rocio C., Carlos Lorenzetti, Ana M., Nelida B., "Genetic Algorithms for Topical Web Search: A Study of Different Mutation Rates", ACM Trans. Inter. Tech., 4(4):378–419, 2005.

[10] Mercy T., Naomie S., "A Framework for Genetic-Based Fusion of Similarity Measures In Chemical Compound Retrieval", International Symposium on Bio-Inspired Computing, Puteri Pan Pacific Hotel Johor Bahru, 5 - 7 September 2005.

[11] Ahmed A. A. Radwan, Bahgat A. Abdel Latef, Abdel Mgeid A. Ali, Osman A. Sadek, "Using Genetic Algorithm to Improve Information Retrieval Systems", proceedings of world academy of since, engineering and technology, volume 17, ISSN 1307-6884, 2006.

[12] Abdelmgeid A., "Applying Genetic Algorithm in Query Improvement Problem", International Journal "Information Technologies and Knowledge, Vol.1, p 309-316. 2007.

[13] Khoja, S., "APT:Arabic part-of-speech tagger", proceedings of the student workshop at second meeting of north American chapter of Association for Copmputational Linguistics (NAACL2001), Pittsburgh, Pennsylvania, pp. 20-26, 2001.

[14] yahaya, A., "on the Complexity of the initial stage of Arabic text processing", First Great Lakes Computer Science Conference, Kalamazoo, Michigan, USA, October, 1989.

[15] Goweder, A., De Roeck, A., "Assessment of a Significant Arabic Corpus", Arabic Natural Language Processing Workshop (ACL2001), Toulouse, France. Downloaded from: (http://www.elsnet.org/acl2001 arabic.html).

[16] I. Hmedi, and G. Kanaan and M. Evens, "design and implementation of automatic indexing for information retrieval with Arabic documents", Journal of American society for information science, Volume 48 Issue 10, pp. 867-881, 1997.

[17] Bassam Al-Shargabi, Islam Amro, and Ghassan Kanaan, "Exploit Genetic Algorithm to Enhance Arabic Information Retrieval", 3rd International Conference on Arabic Language Processing (CITALA’09), Rabat, Morocco, pp. 37-41, 2009.

[18] Fatemeh Dashti, and Solmaz Abdollahi Zad," Optimizing the data search results in web using Genetic Algorithm", international journal of advanced engineering and technologies, Vol 1, Issue No. 1, 016 – 022, ISSN: 2230-781, 2010.

First Author Dr. Eman Fares Al-mashakba is Assistant Professor at the Computer Information Systems department, faculty of computer sciences and informatics, Irbed University, Irbed, Jordan. She holds a Ph.D. in Computer Information Systems from the Arab Academy for Banking & Financial Sciences University, Jordan. His research interests include information retrieval, Security, E-learning and image processing.

Second Author Dr. Feras Fares Al-mashakba is Assistant Professor at the Computer Information Systems department, faculty of computer sciences and informatics, Amman Arab University, Amman, Jordan. He holds a Ph.D. in Computer Information Systems from the Arab Academy for Banking & Financial Sciences University, Jordan. His research interests include information retrieval, artificial intelligence, M-commerce.

Third Author Dr. Mohammad Othman Nassar is Assistant Professor at the Computer Information Systems department, faculty of computer sciences and informatics, Amman Arab University, Amman, Jordan. He worked as Assistant Professor at the Computer Information Systems department in the Arab Academy for Banking & Financial Sciences University. He holds a Ph.D. in Computer Information Systems from the Arab Academy for Banking & Financial Sciences University, Jordan. His research interests include information retrieval, supply chain management, reengineering, outsourcing, and security. Dr. Nassar has published more than 12 articles in these fields in various journals and international conferences. He is included in the Panel of referees of "International Journal of Modeling and Optimization" and in the "International Journal of Computer Theory and Engineering", he was reviewer in the 2011 3rd International Conference on Machine Learning and Computing, also he is currently reviewer in A collection of open access journals called (academic journals).

Weighted Bit Rate Allocation in JPEG2000 Tile Encoding

Singara Singh[1], R. K. Sharma[2] and M. K. Sharma[3]

[1]Assistant Professor, [2]Professor, [3]Associate Professor

School of Mathematics & Computer Applications
Thapar University, Patiala, Punjab, INDIA-147004

Abstract

Equal bit rate is assigned to all tiles of an image when compressed with JPEG2000 standard. This bit rate is selected without taking information contents of the tiles into account. This results into poor performance of JPEG2000 standard for the tiles that have higher complexity. We can improve performance of JPEG2000 by assigning higher bit rates to complex tiles. An entropy based weighted bit rate allocation algorithm is proposed in this paper. Experimentations using the proposed algorithm indicate an improvement of up to 2 dB in Peak Signal to Noise Ratio ($PSNR$) and up to 5.352% improvement in Relative Percentage Improvement (RPI) in $PSNR$ in the JPEG2000 reconstructed images.

Keywords: JPEG2000, $PSNR$, MSE, Entropy, Tiles, RPI.

1. Introduction

JPEG2000 is a state-of-art image and video compression standard. It provides better compression performance and other features like region of interest (*ROI*) coding, quality scalability, transmission scalability etc. as compared to the JPEG image compression standard [1-5]. It allows an image to be divided into rectangular blocks of same size called tiles, before compressing the image. In compression process of JPEG2000 encoder, equal bit rate is assigned to each tile of the image. This assignment is suitable for the images with information contents equally distributed throughout the image. However, tiles of an image may have different complexities. Some of the tiles may have larger texture area while others may have larger smooth area. The quality of a reconstructed image varies a lot if all tiles in the image do not have same complexity. As such, we should include the complexity of a tile while assigning a bit rate to it. This is a known fact that entropy of a complex tile is more than the entropy of smooth tile. So, we have here proposed a method to assign bit rate to a tile based on the weights derived using the entropy of a tile. Using this algorithm, tiles of an image have assigned different compression bit rates. Visual quality of reconstructed image is improved by assigning these bit rates to the tiles of an image.

2. Related Work

It is a well known fact that human beings pay more attention to important areas of an image. Motivated by this, Battiato *et al.* [6] proposed a method for allocating bit rate to different tiles of an image on the basis of index of the information content of each tile. Ardizzone *et al.* [7] proposed an adaptive method to assign more bits to the image regions in which errors are more visible, maintaining the global bit rate unchanged. Effectiveness of their method depends on the accuracy of the region classifier. Liu *et al.* [8] proposed an algorithm using the complexity of a tile and motion activity of the tile. However, their algorithm is applicable to the Motion JPEG2000 video sequences only.

3. Proposed Algorithm and Quality Comparison Parameters

3.1 Overview of JPEG2000 Encoder

JPEG2000 encoder consists of many processes, as shown in Fig.1. It allows image to be divided into tiles if the size of the image is very large or the memory available is low. This process is known as tiling. Each tile of an image is compressed independently. After tiling, discrete wavelet transform (*DWT*) is applied on each tile. *DWT* is a subband transform which transfers image/tile from spatial domain to frequency domain. To achieve efficient lossy and lossless compression within a single encoder, two wavelet transforms are employed. The 5/3 reversible and 9/7 irreversible wavelet transforms are chosen for lossless and lossy compressions respectively.

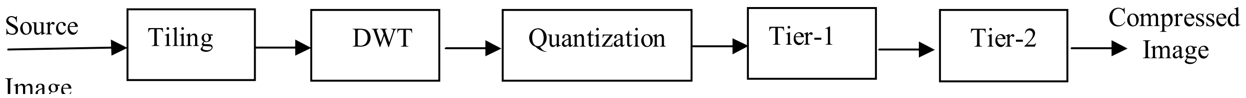


Fig. 1: Block Diagram of JPEG2000 Encoder

After this transform, the wavelet coefficients are quantized to reduce the precision if the lossy compression is chosen. Then wavelet coefficients are entropy encoded by Embedded Block Coding with Optimized Truncation (*EBCOT*) which is a two tier coding algorithm. In *EBCOT*, each wavelet subband is divided into code-blocks. The coefficients of a code-block are represented by their sign-magnitude and encoded from the most significant bit plane to the least significant bit plane by tier-1. Each bit plane is encoded with three coding passes. These passes are significant propagation pass, magnitude refinement pass and cleanup pass. Each pass generates independent bit stream. Finally tier-2 reorders these bit streams into final JPEG2000 output image with rate distortion slope optimized property and the features specified by the user.

3.2 Proposed algorithm

Weighted bit rate allocation method is illustrated in Fig. 2. In this allocation, the weights are derived from the entropy of tiles of an image. This bit rate allocation is passed to the tier-2 process of *EBCOT*, which assigns different bit rate to bit streams of each tile of the source image. After this, the final bit stream is generated by the tier-2 to output the compressed image.

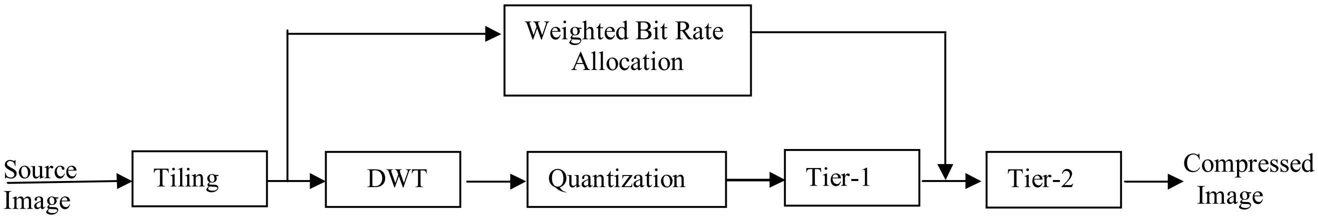


Fig. 2: Block Diagram of JPEG2000 Encoder with weighted bit rate allocation

The entropy *e* of an image is defined as,

$$e = -\sum_{i=1}^{N} P(a_i) \log \left(P(a_i)\right) \qquad (1)$$

where a_i, i = 1, 2, …, N, is the value of i^{th} gray level of original image, N is the total number of different gray levels in the image and $P(a_i)$ is the probability of gray level a_i of the image.

Using this definition, entropy of each tile can be calculated. The weight $w(t)$ assigned to tile t is calculated as,

$$w(t) = \frac{e(t)}{\sum_{t=1}^{N_T} e(t)} \times N_T \qquad (2)$$

where $e(t)$ is entropy of t^{th} tile and N_T is total number of tiles in the image. This can also be noted that,

$$\frac{\sum_{t=1}^{N_T} w(t)}{N_T} = 1 \qquad (3)$$

Number of bits N_b, assigned to a JPEG2000 compressed image is calculated as,

$$N_b = R_0 \times image\_size,$$

where R_0 is the global compression bit rate given by the user and *image_size* is the size of the original image.

Weighted bit rate R_i, based on the entropy, is now assigned to each tile, using the following formula.

$$R_t = R_0 \times w(t) \qquad (4)$$

Thus total number of bits N_b' assigned to the compressed image is given by,

$$\begin{aligned} N_b' &= \sum_{t=1}^{N_T} R_t \times tile\_size \\ &= \sum_{t=1}^{N_T} R_0 \times w(t) \times tile\_size \\ &= R_0 \times N_t \times tile\_size \\ &= R_0 \times image\_size \\ &= N_b \qquad (5) \end{aligned}$$

where *tile_size* is the size of a tile. Eq. (5) shows that total number of bits assigned to the compressed image remains unchanged when the image is compressed using proposed algorithm. The above steps can be summarized in the following algorithm.

Algorithm: Weighted bit rate allocation algorithm
Step 1: Calculate the entropy and weight of each tile of the original image using Eq. (1) and Eq. (2), respectively.

Step 2: Assign weighted bit rate to each tile using Eq. (4). Then compress each tile using JPEG2000 coder.

3.3 Quality Comparison Parameters:

Quality comparison parameters considered in this work are $PSNR$ and RPI in $PSNR$ values. $PSNR$ is determined between the original image and reconstructed image using the following formula.

$$PSNR = 10\,log_{10}\frac{\left(2^{B}-1\right)^{2}}{MSE} \qquad (6)$$

where B is the bit depth of the image and MSE is the mean square error and is defined as,

$$MSE = \sum_{m=1}^{x}\sum_{n=1}^{y}\frac{(A_{mn}-B_{mn})^{2}}{x\times y}$$

where A_{mn} is the pixel of reconstructed image and B_{mn} is the pixel of original image, x and y are the height and width of the images, respectively.

RPI in $PSNR$ is defined as

$$\frac{PSNR_{new}-PSNR_{old}}{PSNR_{old}}\times 100 \qquad (7)$$

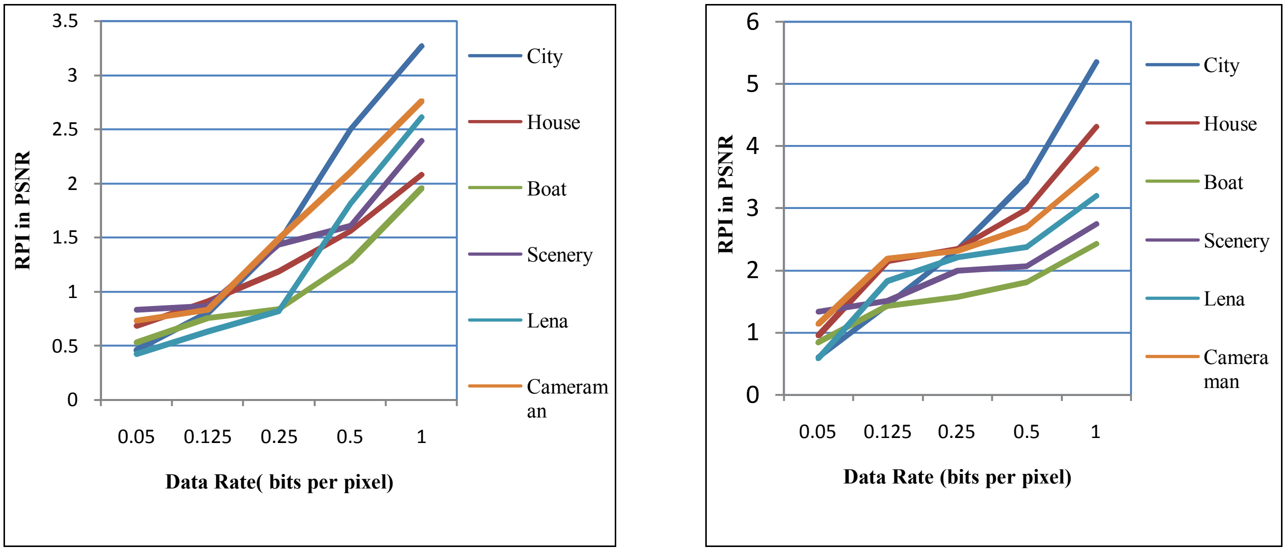
where $PSNR_{new}$ is the $PSNR$ value when proposed algorithm is used with JPEG2000 encoder and $PSNR_{old}$ is the $PSNR$ when existing algorithm is used with JPEG2000 encoder.

4. Results

To implement the proposed algorithm, we modified Kakadu software [9]. In this work, we have considered six standard images taken from literature. These images are compressed using 5 levels wavelet decomposition and a code block size of 64 × 64. In order to demonstrate effectiveness of the proposed algorithm, five bit rates, namely, 1.000, 0.500, 0.250, 0.125 and 0.050 have been considered for each of these six images. The results of this experiment are presented in Table 1.

An analysis of these results is presented in Fig. 3. This contains RPI in $PSNR$ values for images considered in this work for different bit rates. Fig. 3(a) depicts RPI in $PSNR$ values as a function of data rate when the percentage improvement is calculated on the basis of proposed algorithm and algorithm proposed in [7]. Also, Fig. 3(b) depicts this RPI in $PSNR$ values when the percentage improvement is calculated on the basis of proposed algorithm and standard algorithm used in JPEG2000. In Fig. 3(a) $PSNR$ values vary from 1.687% to 3.25% when bit rate is 1.000; vary from 1.246% to 2.560% when bit rate is 0.500; vary from 0.821 % to 1.804% when bit rate is 0.250; vary from 0.696% to1.351 when bit rate is 0.125 and vary from 0.424% to 0.835 % when bit rate is 0.050, as is indicated from Fig. 3(a).

Fig. 3(b) indicates that RPI in $PSNR$ values vary from 2.403% to 5.352% when bit rate is 1.000; vary from 1.815% to 3.4495% when bit rate is 0.500; vary from 1.5693 % to 2.3369% when bit rate is 0.250; vary from 1.4234% to 2.1920 when bit rate is 0.125 and vary from 0.5875% to 1.3359% when bit rate is 0.050 for the six images considered in this work.



(a) (b)

Fig. 3: RPI in $PSNR$ values

Table 1: $PSNR$ comparison of the proposed algorithm with the existing algorithms

Image	Compression Rate (bit per pixels)	$PSNR$ using JPEG2000 Standard	$PSNR$ using algorithm in [7] with JPEG2000 Standard	$PSNR$ using proposed algorithm with JPEG2000 Standard
City	1.000	36.71	37.45	38.73
	0.500	30.94	31.22	32.15
	0.250	27.32	27.57	27.98
	0.125	24.78	24.95	25.16
	0.050	18.79	18.83	18.93
House	1.000	29.55	30.21	30.84
	0.500	27.03	27.41	27.84
	0.250	25.08	25.37	25.68
	0.125	22.44	22.76	22.94
	0.050	19.59	19.81	19.99
Boat	1.000	35.55	35.72	36.43
	0.500	31.69	31.86	32.27
	0.250	28.41	28.61	28.86
	0.125	25.75	25.92	26.15
	0.050	22.29	22.35	22.49
Scenery	1.000	44.74	44.90	45.98
	0.500	40.81	40.99	41.66
	0.250	36.93	37.14	37.67
	0.125	34.07	34.29	34.59
	0.050	20.88	20.99	21.17
Lena	1.000	33.16	33.35	34.23
	0.500	31.21	31.38	31.96
	0.250	28.95	29.35	29.60
	0.125	24.80	25.09	25.25
	0.050	21.02	21.05	21.15
Cameraman	1.000	44.37	44.75	45.99
	0.500	38.08	38.29	39.11
	0.250	32.96	33.24	33.75
	0.125	28.60	28.98	29.23
	0.050	22.75	22.85	23.03

This can also be inferred from Table 1 that $PSNR$ values for all images and for all bit rates is improved when the proposed algorithm is used vis-a-vis the algorithm implemented in JPEG2000 standard and algorithm proposed in [7].

5. Conclusions

In this paper, a weighted bit rate allocation algorithm for JPEG2000 image tiles has been proposed. The proposed methodology improves the $PSNR$ values for all images and for all bit rates considered in this work. It has been observed that the proposed methodology provides better visual quality in JPEG2000 reconstructed images than the conventional approach of JPEG2000 standard. This improvement has been shown taking place when compared with JPEG2000 encoder and also with the algorithm proposed by Ardizzone *et al.*[7].

References

[1] ISO/IEC 15444-1: “Information Technology – JPEG2000 Image Coding System - Part 1: Core Coding System”, 2001

[2] Taubman D. S. and Marcellin M. W. : “JPEG2000: Image Compression Fundamentals, Standards and Practice”, Kluwer, Boston, MA, 2002.

[3] Kang K. J.: “A Fast and Dynamic Region-of-Interest Coding Method Based on the Patterns in JPEG2000 Images”, Int. J. of Innovative Computing, Information and Control, Vol. 5, No. 5, 2009, pp. 1161-1170.

[4] Seo Y. G. and Kang K. J.: “A Slope Information Based Fast Mask Generation Technique for ROI Coding in JPEG2000”, Int. J. of Innovative Computing, Information and Control, Vol. 6, No. 6, 2010, pp. 2817-2826.

[5] Kale V. U., and Deshmukh S. M., “Visually improved image compression by combining EZW encoding with texture modeling using Huffman Encoder”, International Journal of Computer Science Issues, Vol. 7, No. 3, 2010, pp. 28-38 .

[6] Battiato S., Buemi A. Impoco G., and Mancuso M., “JPEG2000 Coded Images Optimization Using a Content Dependent Approach”, IEEE Trans. on Consumer Electronics, Vol. 48, No. 3, 2002, pp. 400-408.

[7] Ardizzone E., Cascia M. L. and Testa F.: “A New Algorithm for Bit Rate Allocation in JPEG2000 Tile Encoding”, IEEE Proc. 12th Int. Conf. on Image Analysis and Processing, Italy, 2003, pp. 658-661.

[8] Liu J., and Zhang D.: “A Novel Bit Rate Allocation Algorithm for Motion JPEG2000”, IEEE Proc. 6th World Congress on Intelligent Control and Automation, China, June 2006, pp. 9907-9910.

[9] Kakadu Software, www.kakadusoftware.com.

Singara Singh is currently working as Assistant Professor in School of Mathematics and Computer Applications, Thapar University, Patiala, India. He received his M. Sc. and M. Tech. degrees in 1998 and 2000, respectively. He is currently pursuing Ph. D. degree in image compression. His research interests include image processing, wireless networks and data security.

Dr. R. K. Sharma is Professor in School of Mathematics and Computer Applications, Thapar University, Patiala, India. He obtained his Ph. D. degree in 1993 from Indian Institute of Technology, Roorkee, India. He has published more than 30 papers in the area of pattern recognition, neural networks and ATM networks. His research interests include soft computing, neural networks, and statistical methods in NLP.

Dr. M. K. Sharma is Associate Professor in School of Mathematics and Computer Applications, Thapar University, Patiala, India. He obtained his Ph. D. degree from Indian Institute of Technology, Roorkee, India. He has published several papers in the area of theoretical astrophysics and operations research. His research interests include theoretical astrophysics, multiobjective optimization and image processing.

WiBro Mobility Simulation Model

[1]Junaid Qayyum, [2]Shahid Latif, [3]Faheem Khan, [4]Muhammad LaL, [5]Shahzad Hameed, [6]Asad Malook

[1,3,4,5] Gandhara University of Sciences Peshawar
[2]Sarhad University of Sciences and Information Technology
Peshawar, 25000, Pakistan

ABSTRACT

WiBro, or Wireless Broadband, is the newest variety of mobile wireless broadband access. WiBro technology is being developed by the Korean Telecoms industry. It is based on the IEEE 802.16e (Mobile WiMax) international standard. Korean based fixed-line operators KT, SK Telecom were the first to get the licenses by the South Korean government to provide WiBro Commercially. Samsung had a demonstration on WiBro Mobile Phones and Systems at the “APEC IT Exhibition 2006”. WiBro is comprised of two phases namely WiBro Phase I and WiBro Phase II. Samsung Electronics has been extensively contributing to Korea’s WiBro (Wireless Broadband) initiative as well as the IEEE 802.16 standards. The WiBro is a specific subset of the 802.16 standards, specially focusing on supporting full mobility of wireless access systems with OFDMA PHY interface. In this work, we have developed a simulation model of the WiBro system consisting of a set of Base Stations and Mobile Subscriber Stations by using the OPNET Modeler. The simulation model has been utilized to evaluate effective MAC layer throughput, resource usage efficiency, QoS class differentiation, and system capacity and performance under various simulation scenarios.

1. INTRODUCTION

Recently, the IEEE has finalized the 802.16d standard [1], which specifies a set of different physical (PHY) and Medium Access Control (MAC) layers of Broadband Wireless Access (BWA) systems, which are Base Station (BS) and Subscriber Station (SS). The technology enables physically distant users to have access to the high-speed broadband wireless service with a relatively inexpensive cost comparing to existing cable and satellite solutions. In addition, the wireless coverage of 802.16 is much wider than that of 802.11 WLAN technologies while providing more bandwidth to users. Attracted by the above benefits, the industry has already been developing and selling commercial 802.16d systems and the market needs start to grow. Moreover, in order to promote the interoperability and compatibility of 802.16 products, numbers of companies organized the WiMax Forum [6] that offers the interoperability test among various products and fosters the development and commercialization of the products. The 802.16 standard is not only for the fixed BWA systems, but also for the mobile BWA systems. The 802.16e standard [2], which is still being developed, provides the amendment for the 802.16d standard and extends the capability of 802.16 technologies to support subscriber systems with mobility. Samsung Electronics has been intensively working in 802.16e standardization and has contributed several important features of

the 802.16e standard. Recently, Samsung has been elected to be a board member of WiMax forum to lead the mobile task group. In Korea, the government and the industry has been working together to enable the wireless broadband service with mobility support during past years. We named the service as Portable Internet previously, and then changed the name to WiBro. We have our own standard [4, 5], which is a subset of 802.16d/e, to encourage and accelerate nation-wide broadband wireless service. The initial draft service will be launched in the end of 2005. Korea will be the first country in the world to deploy the 802.16e service with the full mobility support by public service providers. In this paper, we present a simulation model of Korea's WiBro systems using the OPNET modeler package including the wireless module. The simulation model consists of a set of BSs, mobile SSs, and several traffic handling node objects. Unlike the previous work [7, 8], we have focused on modeling mobility of SSs and hand-over between different cells as SSs move around. We have executed various performance evaluation tasks to validate the correctness of modeling. Some of the simulation results are presented and explained in this paper. Since Samsung Electronics is currently developing the first WiBro systems as a system vendor, the simulation model is extremely helpful to understand the current performance limitation of WiBro specification, to design system architecture and deploy the device components in systems, and to improve and extend existing features before actual development of the systems. In the next section, the detailed simulation model of WiBro systems is described. We briefly explain the current WiBro and 802.16d/e standard we referred to and how the standard features are modeled. Then, we show how we implement the various modeling components by using OPNET modeler v10.5. Several sets of simulation results are then presented to validate and utilize the modeling.

2. WHY WIBRO?

Provide the way to use high speed internet service not only in home but also in outdoor.

- Maximize the spectral efficiency
- Extend the service coverage
- Reduce the cost per bit
- Low power consumption at AT
- Faster handoff

Figure 1 shows a Typical Public WiBro IP Network.

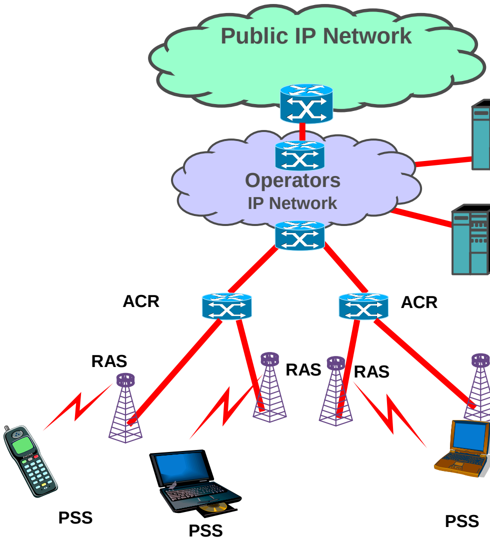


Figure 1

3. BENEFITS FROM IPV6

- Low cost and High Data rate.
- The technology will also offer Quality of Service.
- The spectrum it uses is licensed and correspondingly protected from un-licensed use.
- WiBro can be considered as "Mobile WiMax" which can be used while the receiver is in motion.
- High network capacity

4. WIBRO SYSTEM SPECIFICATION

Korea's Wireless Broadband (WiBro) initiative is pursuing to provide ubiquitous Internet access from various wireless devices with the mobility of up to 60km/h over a distance of several tens of kilometers in the multi-cell environment. It is launched by Korean government and several Korean companies, and the first commercial service will be opened in 2006 by a couple of service providers. The WiBro specification [4,5] released by Korean Telecommunication Technology Association (TTA), is based on a subset of IEEE 802.16 standard. For the radio channel between BS and SS's, Korean government allocates 100MHz frequency bandwidth from 2.3GHz to 2.4GHz. WiBro specifies a communication channel of 9MHz bandwidth, thus, nine individual 9MHz channels are available. Over this radio channel, uplink and downlink access divides a fixed time interval, called frame. Among various physical layer schemes to organize a frame in 802.16, WiBro system only adopts Orthogonal Frequency Division Multiple Access (OFDMA) in Time Division Duplex (TDD) mode. The TDD frame length is 5 msec, and is segmented into the sequence of small fixed-duration logical units, called symbols. The frame structure is fixed as 27 symbols for the downlink subframe and 15 symbols for the uplink subframe. The detailed WiBro frame structure is depicted in Figure 2. (The narrow gaps between subframes are ignored.)

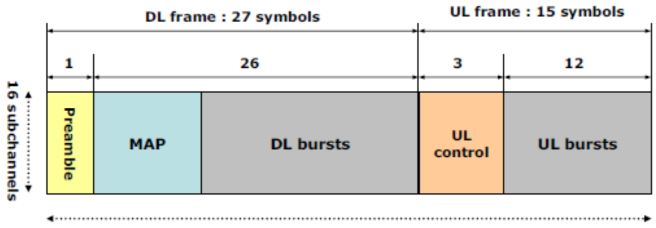
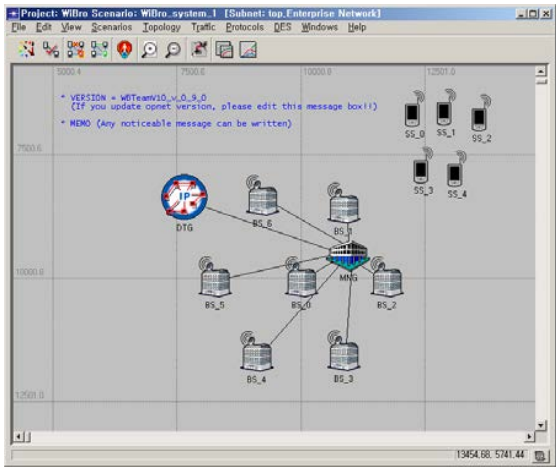


Figure 2: WiBro OFDMA TDD Frame Structure

The encoded data bits are transmitted over a set of subcarriers in wireless communications. In the WiBro specification, there are 864 subcarriers (FFT size is 1024) within 9MHz bandwidth. A set of subcarriers composes a logical transmission unit, called subchannel. While there are several different ways of constructing a subchannel from subcarriers according to the location of subchannels within subframe, the number of subchannels is 16 (in FUSC mode) in the WiBro specification. The unit symbol time of a single subchannel is defined as a slot. This slot is the logical encoding unit of wireless transmission. A bit stream is encoded into a slot, in other words, the bit stream is carried by a subchannel during the period of a symbol time. For downlink, 26 symbol times are available for logical maps and downlink bursts, while 12 symbol times are available for uplink bursts for uplink; 416 slots for downlink and 192 slots for uplink. The MAP consists of Frame Control Header (FCH), DLMAP (downlink map), and UL-MAP (uplink map). The maps guide SS's how to decode the following data bursts. The burst is a set of actual data slots that are allocated by a BS, for either downlink or uplink. SS's are informed when and on which subchannel they need to decode data for downlink, and to encode data for uplink. The BS is responsible for organizing the maps and the bursts in every frame. The 802.16 MAC messages are transferred in each burst. The WiBro specification uses the same MAC message format as 802.16. The MAC message has user payload, 6-byte fixed MAC header, optional 4-byte CRC and optional 12-byte encryption data. There is no difference, when it comes to the functionality and the format of MAC messages, between WiBro and 802.16. In addition to the messages in 802.16d, WiBro adopts standard messages defined in 802.16e for mobility support. Hand-over mechanisms and sleep mode operations are two main features added for the mobility support. The uplink control information subframe is a collection of special-purpose control channels and uses three dedicated symbol times in the uplink subframe. Initial and periodic ranging channels are used for SS's to make network entry and adjustment to wireless channels. Channel Quality Indication (CQI) channels for reporting the previous channel quality and acknowledgement data for Hybrid ARQ (H-ARQ) are included in the control information as well. The uplink transmission is operated in the bandwidth request and grant mechanism. SS's should request a certain amount of bandwidth to BS first when it has some data to transfer, and BS allocates uplink bursts to the SS's after scheduling decisions. There are four different bandwidth allocation service types; Unsolicited Grant Service (UGS), real-time Polling Service (rtPS), non-real-time Polling Service (nrtPS), and Best Effort (BE) service. They are different in how the bandwidth request and grant messages are exchanged between SS and BS. Every uplink session is mapped to one of the service types. The BS is responsible for uplink scheduling as well as downlink scheduling at the same time.

5. SIMULATIONS

OPNET Modeler [9] is a powerful discrete-event simulation tool with easy and convenient development environment and GUI. We used OPNET modeler 10.5 with Wireless Module to develop the simulation models of our WiBro systems. The top-level network browser view of our WiBro simulation model is captured as in Figure 3.



Project: WiBro Scenario: WiBro_system_1 [Subnet: top.Enterprise Network]

* VERSION = WBTeamV10_v_0_9_0
  (If you update opnet version, please edit this message box!!)

* MEMO (Any noticeable message can be written)

Figure 3: WiBro System Model in OPNET Network Browser

There are seven BS nodes connected to MNG node, and the MNG node is connected to DTG node. SS nodes are located separately. The DTG (dynamic traffic generator) node represents traffic source and destination for communicating with SS's. The MNG (management) node represents a centralized router working as backbone network of BS nodes. While the number of SS nodes can be flexibly changed, the number of BS nodes should be fixed to seven to model SS mobility and hand-over between hexagonal shapes of BS cells as in Figure 4.

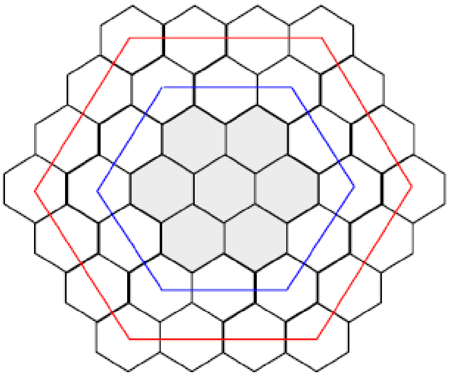


Figure 4: BS Cell Modeling

The seven BS nodes represent seven shaded cells in the center of Figure 4. The radius of a cell is 1000 meters and the initial location of each SS node is randomly given within the shaded area at runtime of simulation. The white cells around the shaded cells are virtual cells to calculate interference and to support wrap-around feature of SS movement. Every shaded cell has six 1st tier neighbor cells and twelve 2nd tier neighbor cells that cause interference. We assume cells farther than 2nd tier cannot add more interference. The SS node is modeled to have one of three different types of mobility as summarized in Table 1. When an SS moves out of a cell, appropriate hand-over steps are performed by the SS and two participating BS's. Instead of using OPNET's trajectory modeling, we implemented our own wrap-around modeling for the movement of SS's to be more general. If an SS moves out of the coverage of shaded cells, we virtually moves SS's going out of one cell to the cell in the opposite direction. For example, if an SS moves from the gray cell of 6 to the white cell of 4, the SS is considered to move from the white cell of 6 to the gray cell of 4. With the cell deployment and mobility modeling, we can generalize our model to be applied to any scenario of WiBro
system configurations.

Table 1: SS Mobility Types

Mobility Types	Speed of SS	Direction Change Scheme (default: at every 30 seconds)
Stationary	0 km/h	No change
Pedestrian	3 km/h	One of 90,180,270,360 degree
Vehicular	60 km/h	One of any degree in 0~360

When the mobile SS travels across multiple BS cells, the handover steps defined in 802.16e are performed at each cell cross. Each mobile SS registers a 'serving BS' when it first enters WiBro network. The serving BS provides network access to the subscribed BS's. When there are multiple BS's sending broadcast messages to SS's, one SS can detect the signal from non-serving BS's and considers them as candidates of 'target BS' by keeping records of the signal power from the candidate BS's. If the signal power of a target BS is stronger than that of the serving BS and maintains the relative strength for a certain period of time (0.3 sec in our simulation), the hand-over steps are performed. The user traffic from/to the moving SS is paused for a short period of hand-over operations, and then resumed. The BS node is modeled as in Figure 4. There are three main processors/queues corresponding to the sublayers in 802.16 standards; convergence sublayer, MAC sublayer, and PHY sublayer. The MAC sublayer contains subqueues to classify uplink and downlink traffic streams by their connection IDs. Radio transmitter/receiver and antenna object from OPNET Wireless Module are used to model the wireless interface to SS's. The SS node is modeled as in Figure 5. In addition to the similar processors/queues in BS node model, there are traffic generator processor and mobility processor in the SS node model. The traffic generator models high-layer applications. We implemented four different application types; VoIP, video streaming, HTTP, and FTP. In order to resemble actual application behaviors over wireless channels, the applications are modeled by 3GPP2's 1xEV-DV application profiles [3]. The traffic generator is flexible to launch any number of application sessions with pre-defined patterns. The mobility processor is for initializing and updating the location of SS periodically during simulation runs. The (x, y) location coordinator of SS is maintained by the mobility processor.

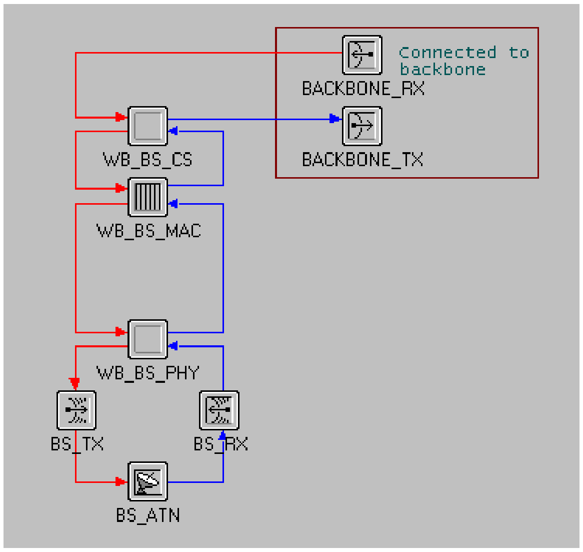


Figure 5: BS Node Model

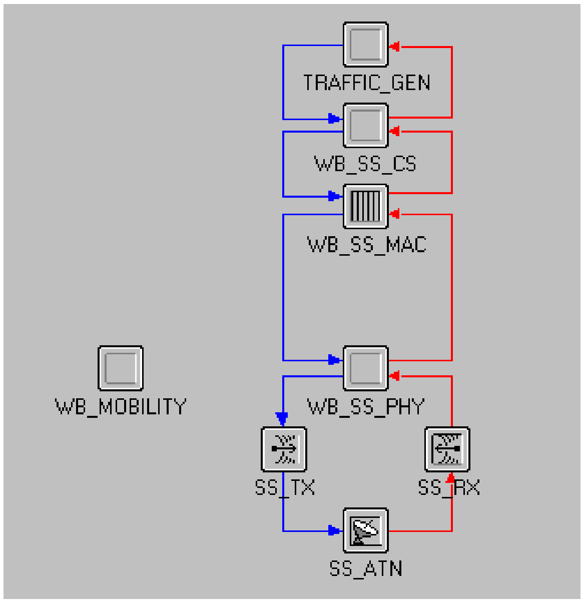


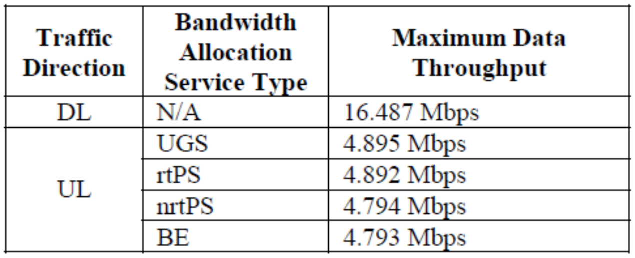
Figure 6: SS Node Model

The physical wireless channel between BS and SS is basically modeled by using OPNET's pipeline stages. However, because the default pipeline stages are modeling only simple TDMA type of wireless channels, we add our own schemes for modeling OFDMA wireless channels. The WiBro system's OFDMA channel has 864 subcarriers of different frequency selection. Since creating 864 individual OPNET wireless channels is totally inefficient, we only virtually model 16 subchannels in WiBro PHY specification within a single OPNET wireless channel. The pathloss fading and shadowing effect is calculated within our own pipeline stages. However, the interference and the modulation schemes of 16 subchannels are separately calculated within PHY processor of BS and SS nodes. In order to accurately model the wireless physical channels, we use results of Samsung's wireless link-level simulations, which are not available to public at this time. Based on the results we set up two relation tables, one for uplink and the other for 4 downlink, which consist of Signal to Interference Ratio (SINR), index of Modulation and Coding Scheme (MCS) level, Packet Error Rate (PER), and speed of mobile SS's. The number of MCS levels for uplink is 8 (from QPSK 1/12 to 16QAM 2/3), and for downlink, the number of MCS levels is 11 (from QPSK 1/12 to 64QAM 5/6). In OPNET simulation, we prepared the tables in OPNET's GDF file format and the tables are loaded at the initial simulation runtime. The GDF tables are looked up twice at each wireless packet communication; one at transmission time and next at receiving time. When in transmission operation, the transmitter needs to determine the MCS level by using (SINR, PER, speed) values. The SINR of previously received burst profile is known by using CQI feedback. The PER is target PER of 0.01, which is given as a global attribute. The speed of SS is fixed and easily obtained. Thus, referring to the appropriate GDF table (either uplink or downlink), the MCS level that satisfies the target PER under the constraints of current SINR and speed of SS is selected.When in receiving operation, the receiver needs to determine the PER by using (SINR, MCS, speed) values. The SINR is calculated by considering various fading effects on the transmitted signal power and interference signal accumulations. The MCS level of burst profiles is written in the profile header information and the speed of SS is fixed. Thus, the GDF table gives PER value of currently received burst profile. The PHY processor drops the burst at the probability of the given PER.

6. SIMULATION RESULTS

In order to obtain the maximum ideal throughput of WiBro systems as reference values, we first assume best conditions of wireless channels; one BS and one fixed SS experience ideal SNR and less than 1% of packet error rate. Thus, they are able to utilize the full capacity of wireless resources by using the best modulation scheme (64QAM 5/6 for downlink and 16QAM 2/3 for uplink) all the time. Table 2 lists the maximum user data (excluding MAC and PHY overhead) throughput values when the overloaded CBR user traffic with the packet size of 1500 bytes is given.

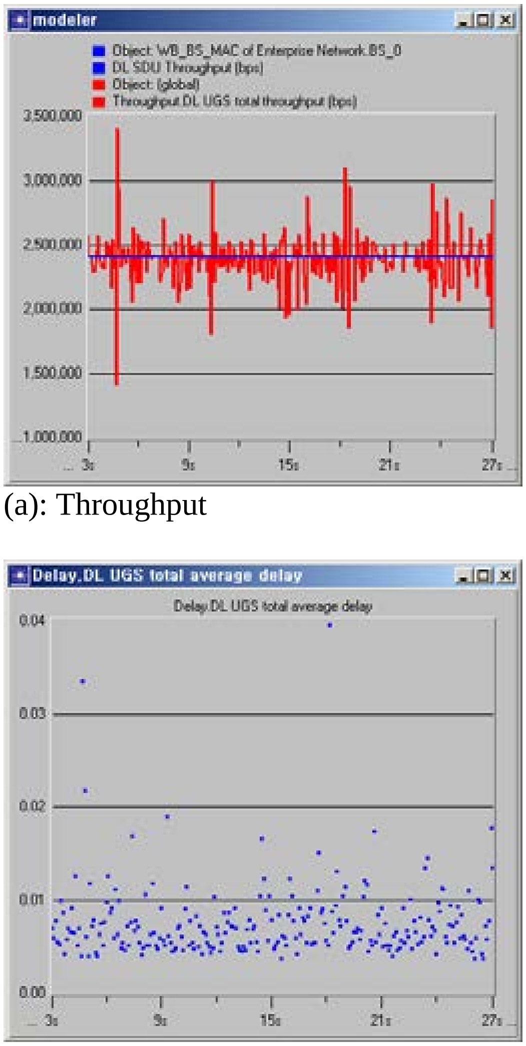
Table 2: Maximum User Data Throughput



Traffic Direction	Bandwidth Allocation Service Type	Maximum Data Throughput
DL	N/A	16.487 Mbps
UL	UGS	4.895 Mbps
	rtPS	4.892 Mbps
	nrtPS	4.794 Mbps
	BE	4.793 Mbps

Due to the asymmetric frame design and different modulation schemes, the maximum throughput for downlink is about 3.37 times more than that for uplink. Among different bandwidth allocation service types in uplink, UGS and rtPS show slightly more throughput than nrtPS and BE because nrtPS and BE enable MAC ARQ in our simulation. The maximum throughput of WiBro is comparatively better than that of symmetric 802.11b WLAN 5.5Mbps theoretically [11] and that of asymmetric CDMA 1x EV-DO (3.1Mbps for downlink and 1.8Mbps for uplink [13]) and WCDMA HSDPA (14Mbps for downlink and 2Mbps uplink [12]). However, in reality, due to the unreliable wireless channel conditions and scheduling overhead, the actual throughput is usually smaller than the ideal maximum throughput. In order to simulate WiBro systems in actual wireless environment, we now enabled all the seven BS's and the wireless channel modeling with various fading and interference. Figure 7 presents an example of downlink CBR traffic over the WiBro wireless channel model. The traffic load is 2.4Mbps with the packet size of 1500 bytes. Because of the dynamic changes of modulation schemes according to the wireless channel conditions, the actual user data throughput received by SS MAC layer fluctuates as in (a) and the packet delay varies up to 40msec (8 frame time delay) as in (b).

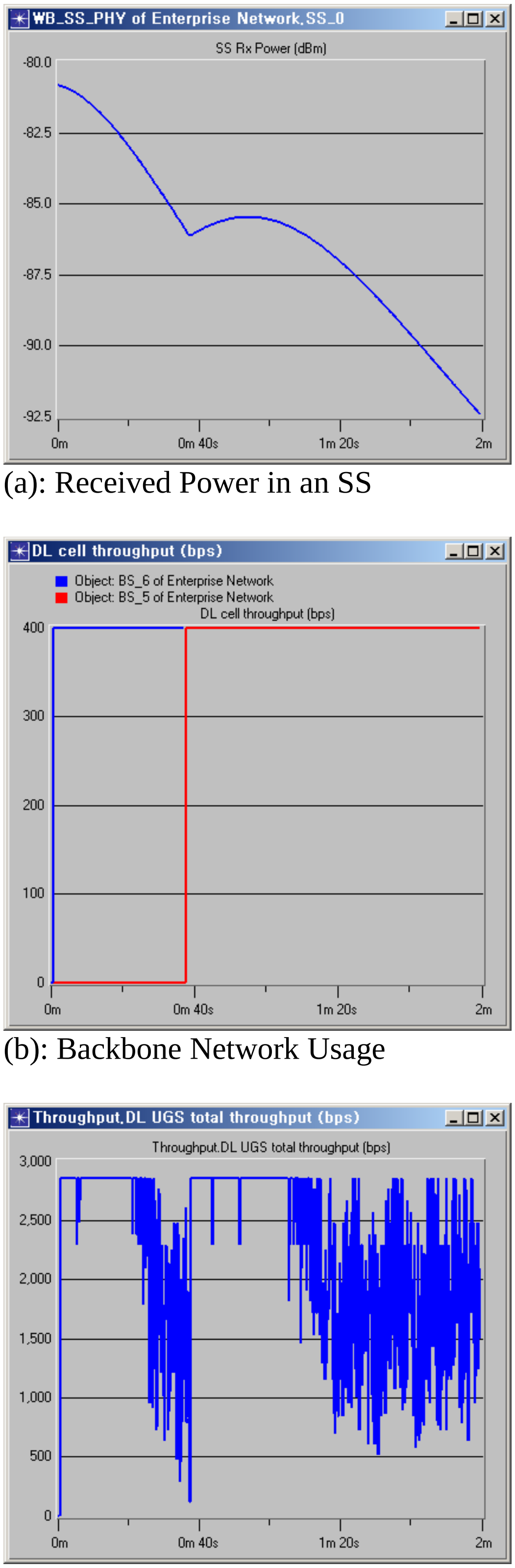
Figure 7: Example Downlink CBR Traffic



(a): Throughput

(b): Packet loss

Next, we verify the hand-over mechanisms of mobile SS when it leaves an existing cell and enters a new cell. Figure 7 shows a typical hand-over example when a hand-over happens at the time of 37.81 sec.



(a): Received Power in an SS

(b): Backbone Network Usage

(c): Actual Downlink User Data Throughput

Figure 8: Example Downlink Hand-over Trace

When a mobile SS travels (60km/h in this example), the received signal power varies as in (a). At the hand-over point, the signal power from a new BS (BS_5) becomes higher than the signal power from the previous BS (BS_6) and the SS decides to move to the new cell. The downlink backbone traffic has been sent to BS_6 is now forwarded to BS_5 as in (b). The actual throughput of the downlink CBR session measured in SS fluctuates around the hand-over time due to increased interference at the cell boundary as in (c).Now we compare the effects of different uplink scheduling service types. Figure 8 shows packet delay of CBR sessions each represents one of four different scheduling services. We configured 100 uplink sessions (25 sessions for each scheduling service type) competing for the limited uplink resources.

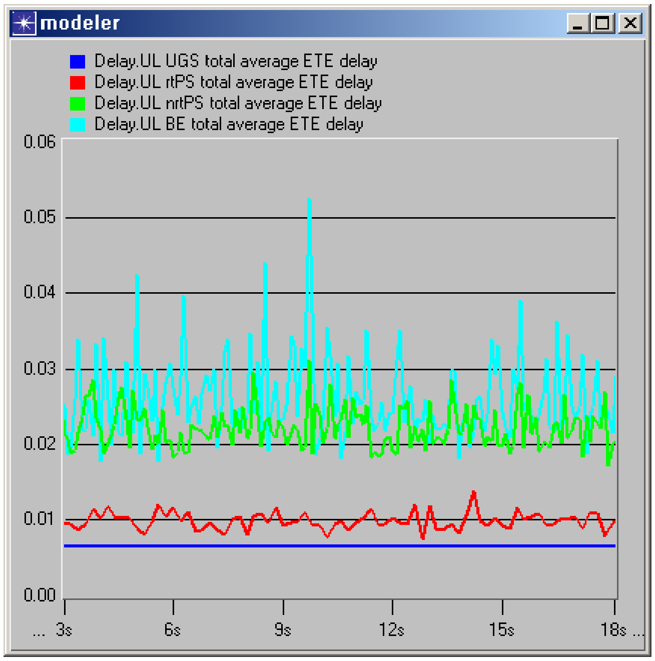


Figure 9: Comparison of UL Polling Services

As depicted, the polling services which guarantee QoS, UGS and rtPS, show smaller and consistent packet delay while two other polling services, nrtPS and BE, show longer and fluctuating packet delay. As the current uplink scheduler uses strict-priority scheduling algorithm (UGS > rtPS > nrtPS > BE) among different polling service types, the above result is straightforward.

7. SIMULATION RESULTS (CAPACITY PLANNING)

First, we measure the average downlink throughput of a BS while increasing the number of mobile SS's having a single downlink CBR application session. The CBR application generates a packet in every 10msec. The wireless channel of a BS is being saturated when we increase the number of SS's. We use the simple round-robin scheduling scheme in the BS scheduler. In order to investigate the impacts of packet size, we use five different packet sizes; 150, 300, 500, 1000, and 1500 bytes. As the inter-arrival time between packets is constant, the traffic load increases proportionally. The average downlink data throughput which is the sum of average data throughput of all SS's stops increasing when the wireless channel reaches to the saturation level as in Figure 10. Comparing to the ideal maximum throughput, 16.487Mbps, in Table 2, we found the maximum throughput under real situations is less than a half of the ideal value. Moreover, we noticed that the saturation level is increasing as the packet size increases. The behavior is explained as more bandwidth is required for MAC headers in the cases with smaller packet sizes because there should be more number of packets to make the wireless channels to be saturated. It is more likely to have more number of downlink packets in each frame for the cases with smaller packet sizes. Note that the size of MAC headers is proportionally increasing when the number of packets increases. The packet delay depicted in Figure 10 shows opposite results. The longer is the packet, the less number of packets are successfully delivered to SS within 20msec delay bound. In the case of 1500-byte packets, only 25% of packets meet the bound when saturated, while 50% of packets survive in the case of 150-byte packets even in the same saturated situation. It is obvious because longer packets are likely to be segmented into several MAC-PDUs to be stored in variable-size DL-bursts and then a packet is successfully delivered only when all the segments are delivered correctly.

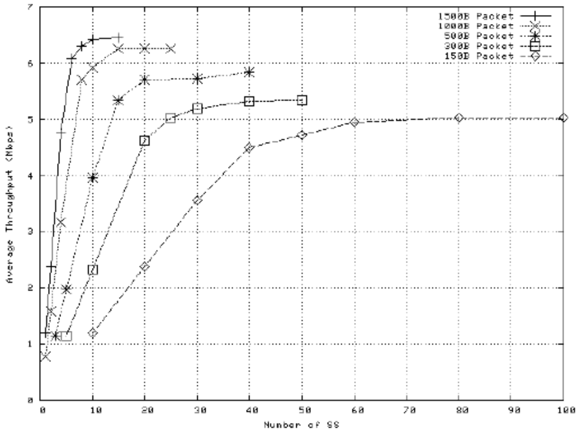


Figure 10: Average Data Throughput of All DL CBR Traffic

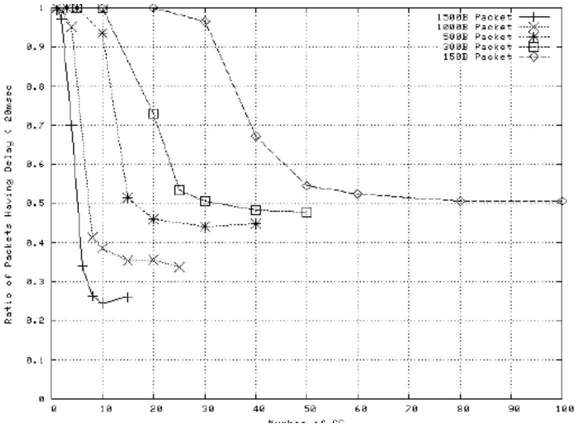


Figure 11: Ratio of Packets Having Less Than 20ms Delay

Next, we set up VoIP application traffic and investigate the capacity of a single BS system when handling numbers of mobile VoIP users in order to perform more realistic analysis. The VoIP application traffic modeling is based on Enhanced Variable Rate CODEC (EVRC) used in CDMA systems. A VoIP session repeats talk-spurt period and silence period, in other words, on period and off period as in Figure 12.

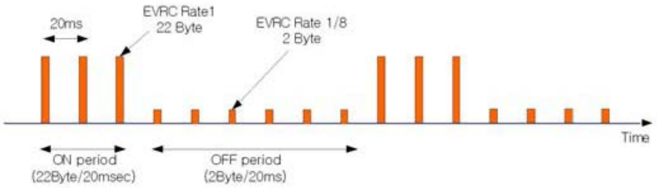


Figure 12: VoIP EVRC Payload Model

Within the on period, 22-byte payload is sent in every 20 msec (EVRC Rate 1). During the off period, only 2-byte payload is sent (EVRC Rate 1/8). The length of the periods follows the exponential distribution; the mean time for the on period is 0.352 sec and 0.650 sec for the off period. In addition to the payload, 40 bytes of IP/TCP/RTP headers are added in uncompressed header mode. In compressed header mode, the header size is compressed to only 4 bytes. While increasing the number of mobile SS's each running VoIP application, we measured the packet delay between BS and SS both in uncompressed header mode and compressed header mode. The measured result is presented in Figure 12. Due to the asymmetry between downlink and uplink capacity, the symmetric VoIP application experiences bottleneck in uplink direction first. The UGS uplink scheduling type is used for the VoIP application and the scheduler in BS performs round-robin scheduling among many SS's.

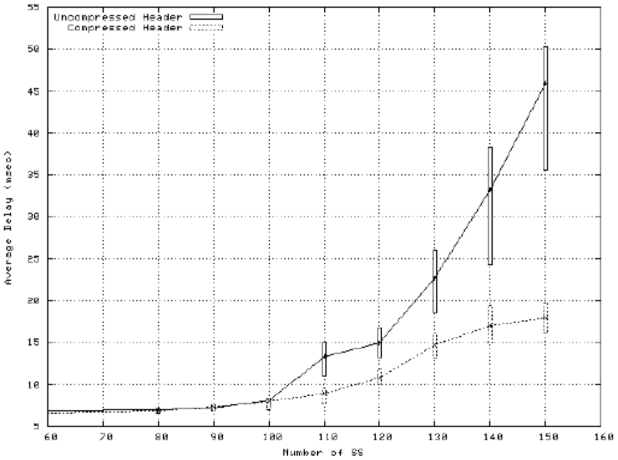


Figure 13: Average VoIP Packet Delay

When the number of SS reaches to 100, the uncompressed VoIP sessions starts experiencing more packet delays and the gaps between average delays of uncompressed VoIP sessions and compressed VoIP sessions becomes wider as the number of VoIP sessions increases. The box represents delay range between 1st quartile and 3rd quartile of each packet delay measurement. As shown, the difference between two quartiles is also increasing as the number of

VoIP sessions increases. With the simulation results, the capacity of WiBro systems is easily projected. If we want packet delay to be less than 20msec, the number of uncompressed VoIP sessions should be controlled to be less than 120. However, we can accept 150 VoIP sessions by using compressed header mode instead.

8. SUMMARY AND FUTURE WORK

In this work we presented an 802.16d/e simulation model named WiBro. The WiBro as well as 802.16d/e systems will not only give wireless users another option of wireless access, but also enable new advanced wireless broadband access by providing much more bandwidth with full-mobility support. As Samsung Electronics plays a leading role in developing WiBro systems and preparing the world-first commercial WiBro service in Korea in 2006, the WiBro simulator is expected to be used in designing actual WiBro system features in many aspects. With OPNET v10.5 modeler package, we implemented the WiBro simulator by modeling OFDMA frame structure, MAC messages, cell architecture, and SS mobility. In order to validate the functionality of the WiBro simulator, we showed several simulation results. First, the ideal maximum throughput of downlink/uplink channel is verified to show the Competitiveness of WiBro standards over existing CDMA, WCDMA or WLAN systems. Then the capacity of actual WiBro systems using mobility modeling of SS's is measured. The maximum capacity of CBR traffic as a general load case and the capacity of VoIP application sessions as a specific real-world example are measured. In both cases, the WiBro simulator gives enough information to understand how much system capacity is used at the given system load. By using this information important system parameters can be analyzed and determined when designing actual systems. Future work includes expanding the simulator to model real WiBro systems Samsung currently develops, enhancing BS scheduler with the proportional fair scheduling algorithm, modifying existing traffic models to adopt OPNET's built-in traffic models as well as TCP/IP stacks, and also verifying mobility support functions which are not fully investigated in this paper.

REFERENCES

[1] IEEE Standard 802.16-2004, “Air Interface for Fixed Broadband Wireless Access Systems,” IEEE, October 2004.

[2] IEEE P802.16e/D7, “Air Interface for Fixed and Mobile Broadband Wireless Access Systems: Amendment for Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands,” IEEE, April 2005.

[3] Third Generation Partnership Project 2 (3GPP2), TSG-C, WG5Evaluation AHG, “1xEV-DV Evaluation Methodology – Addendum (V6),” July 2001.

[4] Korean Telecommunication Technology Association, TTAS.KO-06.0064R1, “Specifications for 2.3GHz Band Portable Internet Service– Physical Layer,” December 2004.

[5] Korean Telecommunication Technology Association, TTAS.KO-06.0065R1, “Specifications for 2.3GHz Band Portable Internet Service– Medium Access Control Layer,” December 2004.

[6] WiMax Forum, http://www.wimaxforum.org.

[7] O.Gusak, N.Oliver, K.Sohraby, “A Simulation Study of 802.16 MAC Layer on a Large-Scale Example,” OPNET Work 2004, August 2004.

[8] S.Ramachandran, C.W.Bostian, S.F.Midkiff, “Performance Evaluation of IEEE 802.16 for Braodband Wireless Access,” OPNET Work 2002, August 2002.

[9] OPNET Modeler, http://www.opnet.com/.

[10] Third Generation Partnership Project 2 (3GPP2), C.R1002-0, “CDMA2000 Evaluation Methodology – Revision 0,” Version 1.0, December 2004.

[11] J.Jun, P.Peddabachagari, M.Sichitiu, “Theoretical Maximum Throughput of IEEE 802.11 and its Applications,” Proc. of the Second IEEE International Symposium on Network Computing and Applications (NCA 2003), pp.249-256, Apr. 2003.

[12] ETSI TS 125 306, “UMTS; UE Radio Access Capabilities Definition (3GPP TS 25.306 Version 6.0.0 Release 6,” December 2003.

[13] Third Generation Partnership Project 2 (3GPP2), TSG-C, C.S0003-A v5.0, “MAC Standard for cdma2000 Spread Spectrum Systems –Revision A,” July 2001.

Implementing and Managing framework for PaaS in Cloud Computing

[1]Junaid Qayyum, [2]Faheem Khan, [3]Muhammad LaL, [4]Fayyaz Gul, [5]Muhammad Sohaib
[6]Fahad Masood
[123456]Gandhara University of Sciences
Peshawar, 25000, Pakistan

ABSTRACT

With the rapid development of Internet and Cloud computing, there are more and more network resources. Sharing, management and on-demand allocation of network resources are particularly important in Cloud computing. Platform as a Service (PaaS) is one of the key services in Cloud computing. PaaS is very attractive for schools, research institutions and enterprises which need reducing IT costs, improving computing platform sharing and meeting license constraints. However, nearly all current available cloud computing platforms are either proprietary or their software infrastructure is invisible to the research community except for a few open-source platforms. For universities and research institutes, more open and testable experimental platforms are needed in a lab-level with PCs. In this paper, a framework for managing PaaS in a virtual Cloud computing lab is developed. The framework implements the user management, resource management and access management. The system has good expandability and can improve resource's sharing and utilization.

1. INTRODUCTION

Cloud computing is developing based on years' achievement on virtualization, Grid computing, Web computing, utility computing and related technologies. Cloud computing provides both platforms and applications on-demand through Internet or intranet [1][2][7][13]. Some examples of emerging Cloud computing platforms are Google App Engine [14], IBM blue Cloud [16], Amazon EC2 [17] and Microsoft Azure [18]. The Cloud allows sharing, allocation and aggregation of software, computational and storage network resources on-demand. Some of the key benefits of Cloud computing include hiding and abstraction of complexity, virtualized resources and efficient use of distributed resources [2]; Cloud computing is still considered in its infancy, there are many challenging issues waiting for tackling [1][2][5][6][7][13]. Platform as a Service (PaaS) is one of the key services in Cloud computing. "PaaS is the delivery of a computing platform and solution stack as a service without software downloads or installation for developers, IT managers or end-users,… It's also known as Cloudware." [14] It is very important to develop an on-demand resource management system for PaaS in Cloud environments. In this paper, a framework for platform as a service is developed. It is also possible to apply the proposed solution to real and vitual Cloud computing environment. The system implements the user management, resource management and remote access. For schools, research institutes and small/medium size enterprises, reducing the IT cost is especially important. For example, in the traditional school lab, because of software license and hardware constraints, many useful application software and platforms are not accessible to students “anytime and anywhere”. This problem may be solved using PaaS in Cloud computing. Through virtualization and other resource sharing mechanisms, Cloud computing can dramatically reduces user costs and meet large-scale applications’ demands. Using virtualization techniques, it is possible to open a few platforms in a single physical machine (Windows, Linux or others) so that resources can be shared better and more users can be served. Most of Cloud computing platform is based on virtualized environments. In a virtualized Cloud computing lab, there are four major parts: software and hardware platforms provided from real and virtualized servers (narrowly speaking, PaaS resources); resource management node; database servers and users who access resources through Internet or Intranet. Generally speaking, above mentioned platforms and users can all be called resources in the Cloud. In the following sections, we consider a framework of design and implementation of PaaS in the Cloud, especially focusing on the resource management. Section 3 discusses the design architecture and major modules in the system; section 4 introduces the implementation technologies and operational environment; Related work in the literature are introduced in Section 5; finally a conclusion is provided in section 6.

2. CLOUD COMPUTING HIERARCHICAL STRUCTURE

The present study achievements haven't achieved an agreement on the definition of "cloud" and "cloud computing". Could computing is generally viewed as the development of Parallel Computing, Distributed Computing and Grid Computing or the commercial realization of these computer science conceptions. Cloud computing is a production of the mixing, evolution and development of several conceptions such like virtualization, utility computing, IaaS, PaaS and SaaS.

2.1 SaaS (Software as-a-Service)

SaaS is the supreme, first appeared and the most common type of cloud computing. It includes a complete application provided to a service through multitenancy demand. The software instances are used as providers' infrastructure and provide services for several end-users or customer organizations. The basic idea of SaaS is to put software on providers' servers and let the operators in charge of the management of maintenance and upgrades. Users who purchase the software only buy the network's permission to use the software instead of installing the software locally. As for the users, they will save the expenses of server and software license. As for the suppliers, they only need to maintain a program so they will reduce the cost.

2.2 PaaS (Platform as-a-Service)

PaaS is not only abstract packages of development environment and also packages of effective service load. PaaS productions can execute the software development and testing of various stages or be used for a certain field. PaaS service can provide great flexibility, but might be affected by the suppliers' ability. Users can develop their own program by middlemen's infrastructure equipments and deliver it through Internet and their server to other users.

2.3 IaaS (Infrastructure as-a-Service)

IaaS is in the lowest level and is a mean of providing basic storage and computing ability on line as a standardize service. Servers, storage systems, switches and routers and other systems are operable and can be used to handle workload from application components to the high performance computing applications.

3. DESIGNING PaaS SYSTEM

3.1 The Architecture of a Virtual Cloud Computing Lab

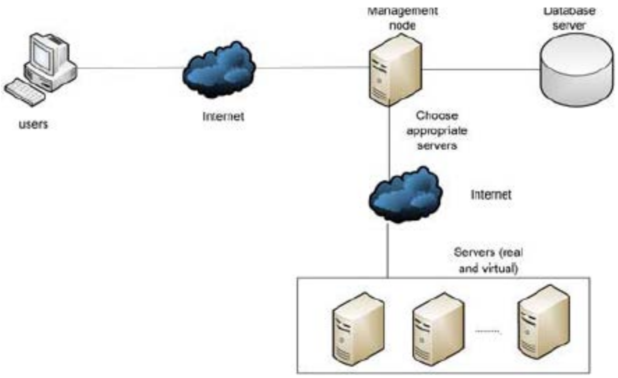
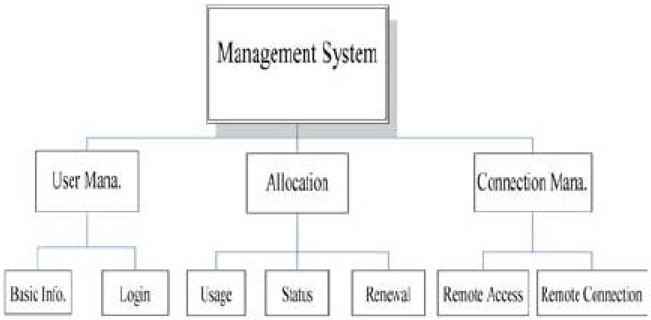


Figure 1 A virtual Cloud Computing Lab

A simplified Cloud computing environment is shown in Figure 1, where users send requests for computing platforms through Internet or intranet (Cloud); management node which may be physically in the same cloud as server groups, verifies the user account, finds available real and virtual servers with requested platforms and allocates them to the user for some periods of time; database servers keeps users authentication, resource availability and other information; after some time, the user finishes the service and leaves the system or chooses to renew. This paper discusses how to design and implement the lab with focus on the management system.

Management System of On-demand Resource Allocation.



The management system includes a user management module, resource allocation module and connection management module. These three modules can be divided into the corresponding sub-modules. User Management module includes basic information management and user access management. Basic information management is mainly concerned with users' information changes to database records; user login management is mainly responsible for the user login and authentication, as well as the user interface. Resource allocation subsystem is the core of the management system, including resource usage, resource status and resource renewal subsystem. Resource usage manages the immediate users and books resources for future users; resources status management maintenances status of all resources; resources renewal management lets user renew the use of resources if possible. Connection management module is to deal with users' accessing resources, including remote access management and remote connection management. These can be done in the remote servers together with management node. PaaS resources can be controlled by one management node or many nodes in the Cloud.

3.2 Communication Among Core Modules

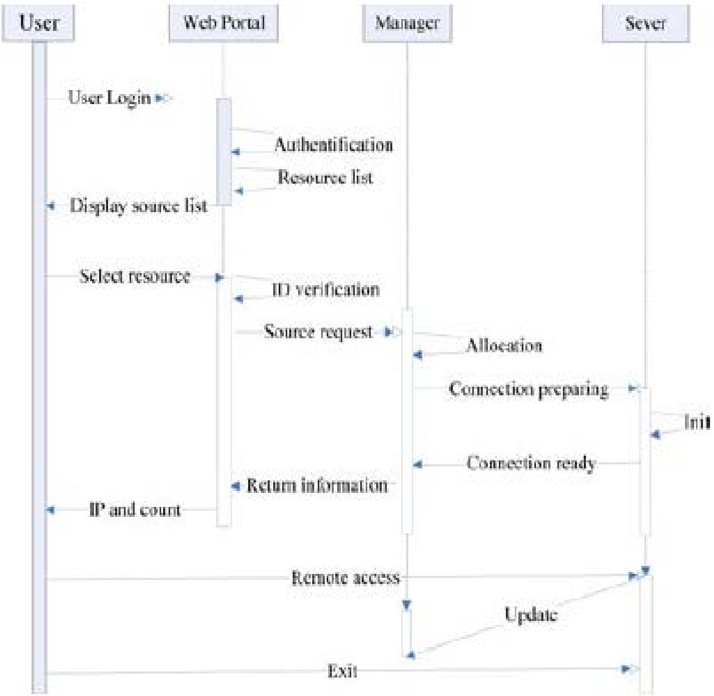


Figure 3 Communication Among Core Modules

In Figure 3, Web Portal is users Web access interface; Manager refers the resource allocation manager; Server refers to a group of real or virtual servers. From the figure we can see that the major communications among core modules: users access Web servers and resources list, and selects resources; Web server forwards the user request to resource management node for processing; then, resource management node sends back Web server the resources information by IP address and users account; Finally, users get access to resources in real or virtual servers. The management system of PaaS needs to coordinate among these four parts to efficiently manage users, platforms resource and remote connections.

3.3 User Management

There may be four kinds of users in PaaS: end users, personnel who manage access to the resources and allocate resources, creators of the PaaS service and PaaS framework developers. In this paper users refer the end-user only, who accesses PaaS service through a web portal. The user can select from a menu list of a combination of applications and operating systems. The user can request for immediate use or for sometime in the future (reservation). There are time windows for user to choose. Once authenticated, user can access remote PaaS service use security remote connection such as openSSH.

3.4 Database Management

Authentication, resource availability and other information is kept in a database server. Therefore, database server has to maintain and manage four kinds of information: user information (UserInfo), platform information (resourceInfo), platform state information (stateInfo) and user connection information (connectionInfo). Their contents and relationship are shown in Figure 4 MySQL is used for this purpose to keep information of authentication, resource availability and other information.

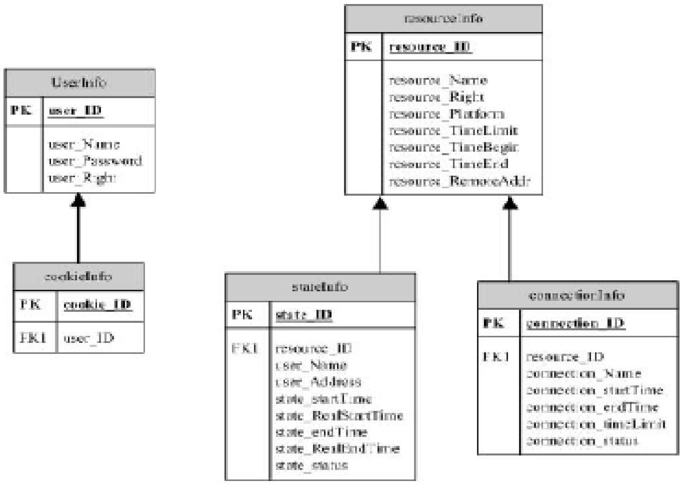


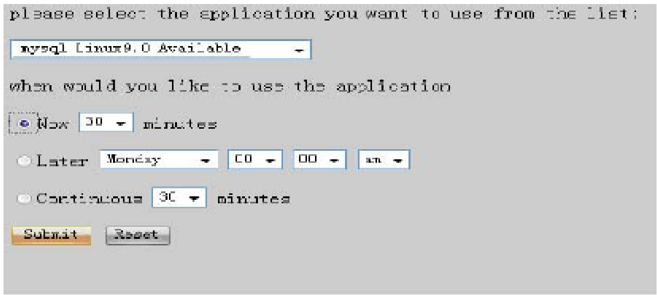
Figure 4 Database Management System

3.5 Virtualization in Operating System Level

Virtualization is one of key technology in Cloud computing. There are many levels of virtualization such as operating system level, hardware location level and network level. Operating system level virtualization is considered only in this paper. Using VMware workstation and other related virtualization software, it is possible to open a few platforms in a single physical machine (Windows, Linux or others) so that resources can be shared Efficiently and more users can be served.

4. IMPLEMENTATIONAL AND OPERATIONAL ENVIRONMENT

The system is developed using open resources including Apache web server, MySQL database server, OpenSSH remote access tools; also VMWare workstation 5.5 is used to create virtual platforms.The user can select appropriate operating platforms with application software. There are two kinds of choices: immediate (now) application and reservation for future use. The user should choose amount of time for his application.



please select the application you want to use from the list:

mysql Linux9.0 Available

when would you like to use the application

● Now 30 minutes

○ Later Monday 00 00 am

○ Continuous 30 minutes

Figure 5 Web Interface for PaaS

The system will be open source in the near future under Eclipse open source license. Theoretically it is possible to provide and manage hundreds of real and virtual platforms; more test and evaluation results are conducting in the following work.

5. RELATED WORK

There may be no consistent definition for Cloud computing yet, however, practitioners are designing and implementing some application examples such as Google App Engine, IBM blue Cloud, Amazon EC2 and Microsoft Azure. There are many pioneering work in this area, many people think that Cloud computing becomes popular after IBM and Google jointly announced Cloud computing plan in 2007. IBM introduces its blue Cloud in [2][16], Google's App Engine[15] and related Google file system [8], BigTable [4] and MapReduce [6] are considered to have laid foundation for Cloud computing. A virtual computing lab (and then Cloud computing) was built since 2004 [12]. Cloud implementation and research related issues are discussed in [2][7][12][13]. As this writing, more than 30,000 teachers and students use VCL [12] at NCSU each year. Eucalyptus [7] is among one of a few an open-source systems for implementing on-premise private and hybrid clouds using the hardware and software infrastructure. Eucalyptus adds capabilities such as end-user customization, self-service provisioning, and legacy application support to data center virtualization features, making IT customer service easier, more fully featured, and less expensive. To understand Cloud computing better and quantify the performance of scheduling and allocation policy on a Cloud infrastructure, simulation tool CloudSim is proposed [3]. Approaches of dimensioning a virtual computing lab with job priorities and QoS constraints is discussed in [9]. Three techniques to improve the efficiency of virtual Cloud computing lab based on queuing model are introduced in [10]; some of these techniques are applied in this paper. Adaptive dimensioning approaches of Cloud datacenters are introduced in [11]. There are many other related work and many more to come in Cloud computing.

6. CONCLUSION

In this paper, a framework implementing and managing platform as a service in a virtual Cloud computing lab is developed. The system has good expandability and can improve resource's sharing and utilization. In the future we will extend the framework to include imaging of software and hardware platforms, load balancing and complete automatic provisioning of resources so that the system can be applied in large-scale and distributed environment.

REFERENCES

[1] Armbrust, M., et al.: Above the Clouds: A Berkeley View of Cloud Computing, Tech. Reprot No. UCB/EECS-2009-28, 2009.

[2] Boss, G., et al.: Cloud Computing, IBM Corporation white paper, Oct. 2007.

[3] Calheiros, R. N. , et al. : Cloudsim: A Novel Framework for Modeling and Simulation of Cloud Computing Infrastructures and Services, Technical Report, GRIDS-TR-2009-1, Grid Computing and Distributed Systems Laboratory, The University of Melbourne, Australia,
March 13, 2009.

[4] Chang, F. et al., Bigtable: A Distributed Storage System for Structured Data, in the proceedings of OSDI 2006.

[5] Chen, K., and Zheng, WM. , Cloud Computing: System Instances and Current Research, Journal of Software, Vol.20, No.5, May 2009, pp.1337□1348.

[6] Dean, J., And Ghemawat, S. , MapReduce: Simplied data processing on large clusters. In Proc. of the 6th OSDI(Dec. 2004), pp. 137.150.

[7] Daniel, N., et al., The Eucalyptus Open-source Cloud-computing System, in Proceedings of 9th IEEE International Symposium on Cluster Computing and the Grid, Shanghai, China, 2008.

[8] Ghemawat, S., Gobioff, H., and Leung, S.-T. The Google file system. In Proc. of the 19th ACM SOSP (Dec.2003), pp. 29.43.

[9] Tian, WH. and Perros, H. G., Dimensioning a Virtual Computing Lab with Job Priorities and QoS Constraints, In the proceedings of 2nd International Conference on the Virtual Computing Initiative ,
pp.103-110, May 2008, Research Triangle Park, IBM headquarter, NC, USA.

[10] Tian, WH., Three Ways to Improve the Efficiency of Virtual/Clould Computing Lab. In the proceedings of The IEEE International Conference on Apperceiving Computing and Intelligence Analysis 2008 (ICACIA'08), Dec. 2008.

[11] Tian, WH., Adaptive Dimensioning of Cloud Datacenters, accepted for publication in the proccedings of The 8th International Conference on Dependable, Autonomic and Secure Computing (DASC-09), Chengdu, China, December 12-14, 2009.

[12] Vouk, Mladen., et al., "Powered by VCL" - Using Virtual Computing Laboratory (VCL) Technology to Power Cloud Computing, Published in the Prelim. Proceedings of the 2nd International
Conference on Virtual Computing Initiative, 15-16 May 2008, RTP, NC, pp. 1-10

[13] Vouk, Mladen A., Cloud Computing – Issues, Research andImplementations,ITI08, pp.23-26-31, June, 2008.7,

[14] wiki, http://en.wikipedia.org/wiki/Platform_as_a_service

[15] Google App Engine, http://code.google.com/intl/zh-CN/appengine/

[16] IBM blue cloud, http://www.ibm.com/grid/

[17] Amazon EC2,http://aws.amazon.com/ec2/

[18]Microsoft-Azure,http://www.microsoft.com/windowsazure/windowsazure

Identifying Reference Objects by Hierarchical Clustering in Java Environment

Rahul Saha , Dr. G. Geetha

Department of Computer Science and Engineering, Lovely Professional University

Phagwara, Punjab, India

Department of Computer Sciences and Applications, Lovely Professional University

Phagwara, Punjab, India

Abstract

Recently Java programming environment has become so popular. Java programming language is a language that is designed to be portable enough to be executed in wide range of computers ranging from cell phones to supercomputers. Computer programs written in Java are compiled into Java Byte code instructions that are suitable for execution by a Java Virtual Machine implementation. Java virtual Machine is commonly implemented in software by means of an interpreter for the Java Virtual Machine instruction set. As an object oriented language, Java utilizes the concept of objects. Our idea is to identify the candidate objects' references in a Java environment through hierarchical cluster analysis using reference stack and execution stack.

Keywords: *Proximity Matrix, Reference Stack, Execution Stack, Euclidean Distance, Object Reference, Dendogram*

1. Introduction

Candidate Objects are those which can be selected as for the options for the objects in a object oriented paradigm. Object identification is a reverse-engineering technique that is largely used to assist the software migration from procedural paradigm to object-oriented paradigm. Object identification facilitates acquiring a precise knowledge of the data items in a program. Object identification reduces the degradation of original design. Object identification typically aims at finding match-up of legacy software components: data structures, and functions, for later building them as object-oriented classes. However, large application consists of numerous data structures and functions; it needs a statistical method to facilitate information classification.

2. Existing Concept

In the paper [1], the authors have described an approach of hierarchical clustering in a procedural language environment using stack and queues. Here the basic functions of a stack and queue are taken to create proximity matrix and pattern matrix. Pattern matrix represents a property set of data (scores or measurements) in a table. Each row stands for a set of properties (a pattern). Proximity matrix represents an index of association (proximity) between pair of patterns. The index can be either similarity index or dissimilarity index and can be computed. by several ways, for example, Simple matching coefficient, Jaccard coefficient, Euclidean distance, Manhattan distance etc. Then they have calculated Euclidean distances between each pair and least values are classified in a cluster. This process goes on until all the properties are successfully clustered. The extracted functions are shown in Table 1 and the relation definitions are shown in Table 2.

Table.1 : Extracted functions

Software component
struct stack
struct queue
struct stack * initStack (int size)
struct queue *initQ ( )
int isEmptyStack(struct stack * s)
int isEmptyQ (struct queue * q)
void push (struct stack * s, int i)
void enQ (struct queue * q, int i)
int pop (struct stack * s)
int deQ (struct queue * q)

Table.2: Relation Definition

Name	**Definition**
R0	Return type is struct stack
R1	Return type is struct queue
R2	Has argument of type struct stack
R3	Has argument of type struct queue
R4	Use field of struct stack
R5	Use field of struct queue

Now in Table 3 the extracted functions and the related definitions are shown. A ‘ X’ mark is put in the cells for each corresponding function and relation definition. In Table 4 the cells are assigned by the values 0 or 1 depending upon the ‘X’ marks of Table 3; 1 is assigned for a ‘X’ mark else 0.

Table.3: Properties in modular case

	R0	**R1**	**R2**	**R3**	**R4**	**R5**
initStack	X				X	
initQ		X				X
isEmptyStack			X		X	
Push			X		X	
enQ				X		X
Pop			X		X	
deQ				X		X

Table. 4: Pattern matrix for modular case

	R0	**R1**	**R2**	**R3**	**R4**	**R5**
initStack	1	0	0	0	1	0
initQ	0	1	0	0	0	1
isEmptyStack	0	0	1	0	1	0
isEmptyQ	0	0	0	1	0	1
Push	0	0	1	0	1	0
enQ	0	0	0	1	0	1
Pop	0	0	1	0	1	0
deQ	0	0	0	1	0	1

Now a proximity matrix is generated using Euclidean-distance method formula given below:

$$d(i,k) = \{ \sum_{j=1}^{t} (x_{ij} - x_{kj})^{2} \}^{1/2} \quad \text{………………(Equation 1)}$$

where, x_{ij} represents the j-ordered attribute of pattern i, x_{kj}

represents the j-ordered attribute of pattern k, and *t* represents the total attribute of pattern. Depending on this proximity matrix clusters are made further by agglomerative method.

3. Our Approach

The authors have illustrated the above in case of procedural language. Our idea is to convert it into object orientation (Java environment).We shall use the same approach but using two stacks i.e. execution stack and reference stack so that the basic approach of the stacks as we have seen in the existing scenario above will be similar and we can integrate it with our Java environment. Table 5 and 6 show the relation definition and extracted functionalities used for our approach.

Table.5: Relation Definition of our approach

Name	**Definition**
R0	Return type is struct execstack
R1	Return type is struct refstack
R2	Has argument of type struct execstack
R3	Has argument of type struct refstack
R4	Use field of struct execstack
R5	Use field of struct refstack

Now, in Table 6 we have extracted some of the functionalities related to the stacks used in our approach that is the functionalities regarding reference stack and execution stack.

Table.6: Extracted functionalities for our approach

Software component
struct execstack
struct refstack
struct execstack * initExec (int size)
struct refstack *initRef (int size)
int isEmptyExec(struct execstack * es)
int isEmptyRef (struct refstack* rs)
void ePush (struct execstack * es, int i)
void rPush (struct refstack * rs, int i)
int ePop (struct execstack * es)
int rPop (struct refstack * rs)
struct execstack* traExec (struct execstack * es)
struct refstack * traRef (struct refstack* rs)

We can use a flowchart diagram to show that how reference and execution stacks are used in identifying the reference to objects. The diagram is given in Fig. 1.

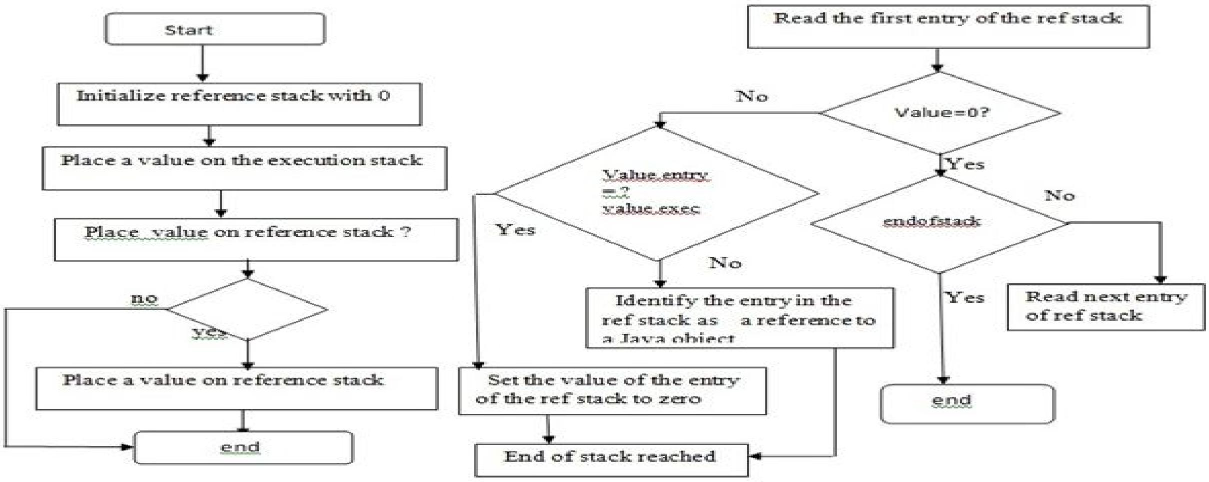

Start
Initialize reference stack with 0
Place a value on the execution stack
Place value on reference stack ?
no
yes
Place a value on reference stack
end
Read the first entry of the ref stack
No
Value=0?
Yes
Value.entry =? value.exec
Yes
No
Identify the entry in the ref stack as a reference to a Java object
Set the value of the entry of the ref stack to zero
End of stack reached
endofstack
No
Yes
Read next entry of ref stack
end

Fig. 1 Flowchart diagram of identifying reference objects

4. Analysis

The relation definition matrix is shown in table 7. We have also generated the pattern matrix which is shown in Table 8.

Table.7: Relation definition matrix for our approach

	R0	R1	R2	R3	R4	R5
initRef	X				X	
initExec		X				X
isEmptyRef				X		X
isEmptyExec			X		X	
ePush			X		X	
rPush				X		X
ePop			X		X	
rPop				X		X
traRef		X		X		X
traExec	X		X		X	

Table.8: Pattern Matrix for our approach

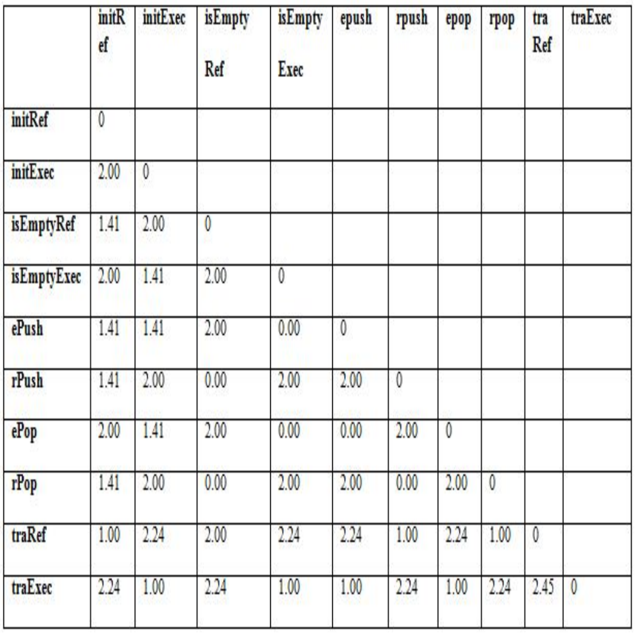
	R0	R1	R2	R3	R4	R5
initRef	0	1	0	0	0	1
initExec	1	0	0	0	1	0
isEmptyRef	0	0	0	1	0	1
isEmptyExec	0	0	1	0	1	0
ePush	0	0	1	0	1	0
rPush	0	0	0	1	0	1
ePop	0	0	1	0	1	0
rPop	0	0	0	1	0	1
traRef	0	1	0	1	0	1
traExec	1	0	1	0	1	0

Now to generate the proximity matrices in each iteration we have used Euclidean formula as defined in Equation 1 earlier. The formula is as below once again:

$$d(i,k) = \{ \sum_{j=1}^{t} (x_{ij} - x_{kj})^{2} \}^{1/2}$$

where, $\mathbf{x_{ij}}$ represents the j-ordered attribute of pattern i, $\mathbf{x_{kj}}$ represents the j-ordered attribute of pattern k**,** and **t** represents the total attribute of pattern. First proximity matrix is shown in the Table 9.

Table.9: First proximity matrix



	initRef	initExec	isEmpty Ref	isEmpty Exec	epush	rpush	epop	rpop	tra Ref	traExec
initRef	0									
initExec	2.00	0								
isEmptyRef	1.41	2.00	0							
isEmptyExec	2.00	1.41	2.00	0						
ePush	1.41	1.41	2.00	0.00	0					
rPush	1.41	2.00	0.00	2.00	2.00	0				
ePop	2.00	1.41	2.00	0.00	0.00	2.00	0			
rPop	1.41	2.00	0.00	2.00	2.00	0.00	2.00	0		
traRef	1.00	2.24	2.00	2.24	2.24	1.00	2.24	1.00	0	
traExec	2.24	1.00	2.24	1.00	1.00	2.24	1.00	2.24	2.45	0

We have considered the least distant (0.00) values from the Table 9 first to form the first round clusters and applied Single linkage rule to form the second proximity matrix in Table 10. The formula for the single linkage rule goes thus:

d [(k), (i,j)] = min {d [(k),(i)], d [(k),(j)] }

where:

d [(k),(i)] represents the similarity between cluster k and cluster i

d [(k),j)] represents the similarity between cluster k and cluster j

d [(k),(i, j)] represents the similarity between cluster k and the newly formed cluster i, j.

Table.10: Second Proximity Matrix

	initRef	initExec	C1	C2	traRef	traExec
initRef	0					
initExec	2.00	0				
C1	1.41	2.00	0			
C2	1.41	1.41	2.00	0		
traRef	1.00	2.24	1.00	2.24	0	
traExec	2.24	1.00	2.24	1.00	2.45	0

C1 = isEmptyRef + rPush + rPop
C2 = isEmptyExec + ePush + ePop

We now have considered the least distant (1.00) values from the Table 10 first to form the next round of clusters and applied the above said Single linkage rule to form the third proximity matrix given in Table 11.

Table.11: Third Proximity Matrix

	C1	C2	C3	C4
C1	0			
C2	2.00	0		
C3	1.00	1.41	0	
C4	2.00	1.00	2.00	0

C3 = traRef + initRef C4 = traExec + initExec

We now have considered the least distant (1.00) values from the Table 11 to form the next round of clusters and applied Single linkage rule to form the fourth proximity matrix given in Table 12.

Table. 12 Fourth Proximity Matrix

	C5	C6
C5	0	
C6	1.41	0

C5=C1 + C3 C6=C2+ C4

We now have considered the least distant (0.00) values from the Table 12 to form the next round of clusters and applied Single linkage rule to form the fifth proximity matrix shown in Table 13.

Table.13: Fifth Proximity Matrix

	C7
C7	0

C7 = C5 + C6

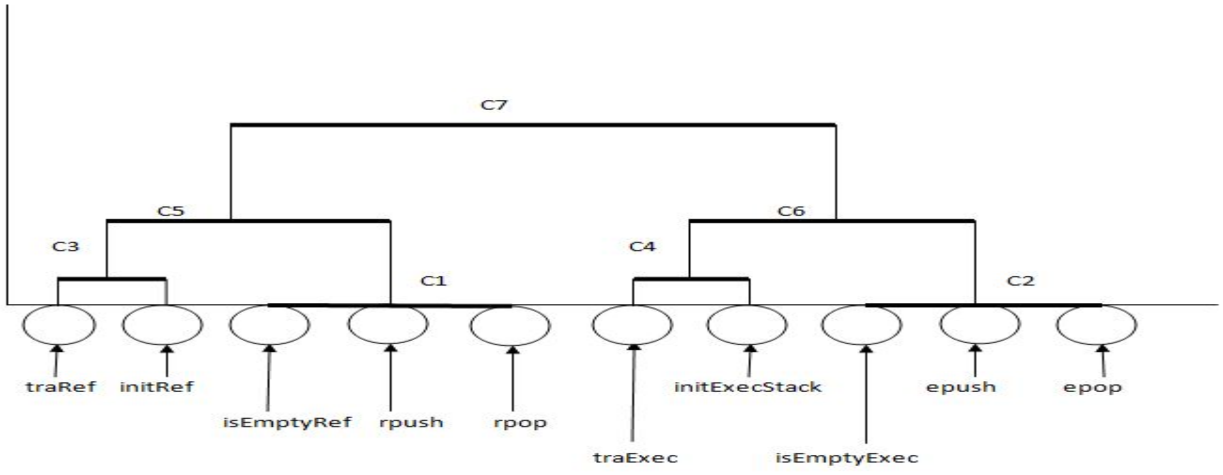


Fig. 2 Dendogram of clusters

4.1.Dendogram

Output of clustering analysis can be represented in various forms depending on the objective of data classification. Output of hierarchical clustering analysis is usually represented in a special type of a tree structure called Dendogram. Our output of clusterization is also represented in Dendogram shown in Fig. 2 above.

5. Conclusion

From the above Dendogram, we can see that the cluster C5 contains all the functionalities that deal with reference stack and cluster C6 contains all the functionalities that deal with execution stack. Cluster C5 can be further divided into C3 and C1 where C3 cluster consists of the functionalities of traversing the reference stack i.e. traRef ( ) and initialization of reference stack i.e. initRef ( ). Cluster C1 consists of the functionalities like to check if the reference stack is empty [isEmptyRef ( )], to insert objects in the reference stack [rpush ( )] , and to delete an object reference [rpop( )].

Similarly, cluster C6 can be further divided into cluster C4 and cluster C2 where C4 consists of traExec ( ) [traversing the execution stack], initExec ( ) [initialization of execution stack] and cluster C2 consists of isEmptyExec ( ), epush( ) and epop( ).

References

[1] Somsak Phattarsukol and Pornsiri Muenchaisri, Identifying Candidate Objects Using Hierarchical Clustering Analysis published in APSEC '01 Proceedings of the Eighth Asia-Pacific on Software Engineering Conference, 2001.

[2] Stepan Sokolov and David Wallman, Identifying and Tracking Object References in a Java programming environment, published in United States Patent, Patent no: US 6804681B2 in Oct 12, 2004.

[3] H.A. Sahraoui, W. Melo, H. Lounis, F. Dumont, Applying Concept Formation Methods To Object Identification In Procedural Code, In Roc. of 12th Conference on Automated Software Engineering, pp. 210 - 218, 1997.

[4] C. Lindig, G. Snelting: Assessing Modular Structure of Legacy Code Based on Mathematical Concept Analysis. Proc. International Conference on Software Engineering (ICSE 97), Boston, USA, May 1997, pp. 349-359.

[5] J. Martin, J. Odell: Object-Oriented Analysis and Design, Prentice Hall 1992.

[6] R. Wirfs-Brock, B. Wilkerson, L. Wiener: Designing Object-Oriented Software, Prentice Hall 1990.

Authors' profile

Rahul Saha is pursuing his M.Tech from Lovely Professional University, Punjab, India in the department of Computer Science and Engineering. His research interest includes Software Engineering, Network Security.

Dr. G. Geetha is the Dean of School of Computer Sciences and Applications. Her research interest includes Cryptography and Software Engineering. She has published more than 30 papers in refereed Journals and Conferences. She is also the Editorial Board of IJACM and IJCRYPTO. She is presently the President of Advanced Computing Research Society.

Development of MIL-STD-1553B Synthesizable IP Core for Avionic Applications

Enumala Srikrishna
Assistant professor, Thandra Paparaya Institute of Science and Technology,
Bobbili, India

L.MadanMohan
Sr.Engineer, Adept Chips

A.Mallikarjuna Prasad
Associate Professor, JNTUK, Kakinada

Abstract

MIL-STD-1553, Digital Time Division Command/Response Multiplex Data Bus, is a military standard (presently in revision B), which has become one of the basic tools being used today for integration of weapon systems. The standard describes the method of communication and the electrical interface requirements for subsystems connected to the data bus. The 1 Mbps serial communication bus is used to achieve aircraft avionic (MIL-STD-1553B) and stores management (MILSTD-1760B) integration. The standard defines four hardware elements. These are 1) The transmission media, 2) Remote terminals, 3) Bus controllers, 4) Bus monitors.

The main objective of this paper is to develop an IP (Intellectual Property) core for the MIL-STD-1553 IC. This IP core can be used as bus monitors or remote terminals or bus monitors. The main advantage of this IP core is to provide small foot print, flexibility and reduce the cost of the system, as we can integrate this with other logic.

1. Introduction

MIL-STD-1553B defines a serial, time division, multiplex command/response data bus. I.e. information is transferred by transmitting a series of data bits one after another; a single transmission path is shared between a number of users by the allocation of time to each user; all transactions take place in response to a command from a single controller. Bus topology is implemented using twisted pair transmission line terminated at each end in its characteristic impedance. Connections are made to the bus via stubs. The standard defines the characteristics of the bus cable, its termination and gives two alternative methods of connecting stubs to the bus 1) Direct coupled and 2) Transformer coupled.

The standard defines three types of terminals. They are bus controller, remote terminal and bus monitor. Bus controller has overall control of all bus activity. All transmissions over the bus are initiated by a command from the bus Controller

The number of remote terminals connected to the bus can be 31 in the range 0-30. Each remote terminal has a unique address. The remote terminals continuously monitor the bus and it receives the message which carries its address. These messages contain the actions to be performed by the remote terminal. The bus monitor monitors the bus and information obtained may be used for offline applications or as a backup. The standard provides the use of two bus for redundancy. The sample bus architecture is given in figure 1.

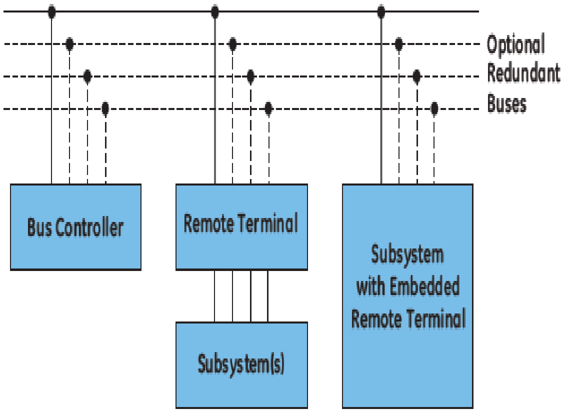


Fig 1:Sample data bus architecture.

The standard provides three types of words and ten types of word formats. Three types of words are:

a) Command word
b) Data word
c) Status word

Command word is issued by bus controller to the remote terminals. Data word can be transmitted by bus controller or remote terminals. The status word is transmitted by the remote terminal to the bus controller in response to the command word. Figure 2 shows different types of words.

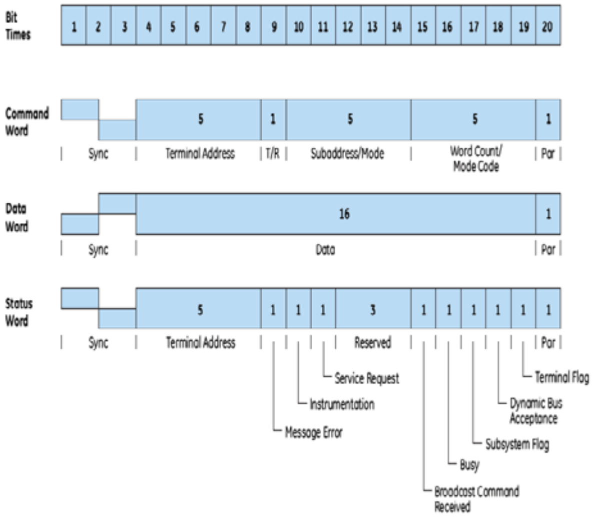


Fig 2 : Types of words.

The message formats can be sub divided into two types.

a) Non-broadcast message formats.
b) Broadcast message formats.

Non broadcast message formats are Bus controller to remote terminal, remote terminal to bus controller, remote terminal to remote terminal, mode command word with data word (transmit), mode command word with data word (receive), mode command without data word. Figure 3 shows different types of non broadcast message formats.

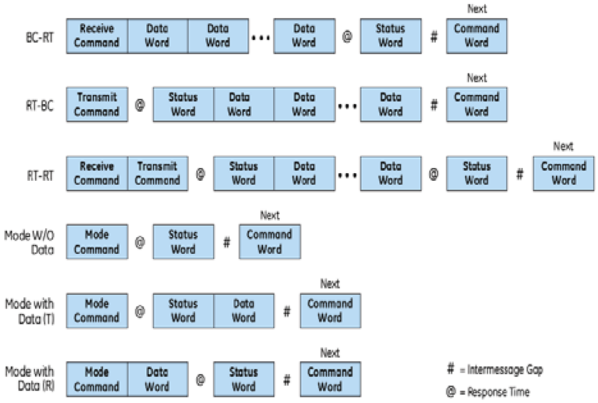


Fig 3: Non broadcast message formats.

Broadcast formats are bus controller to remote terminal transfer; remote terminal to remote terminal transfer, mode command without data word and mode command with data word (receive). Figure 4 shows broadcast message formats

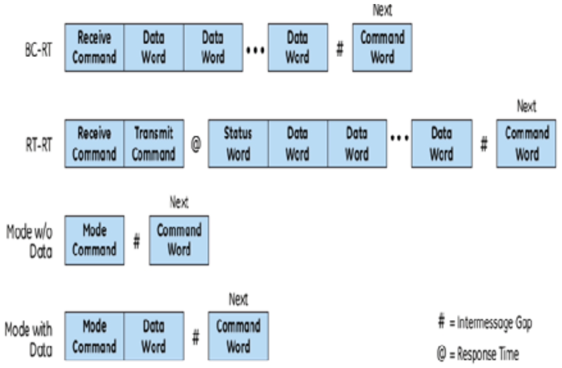


Fig 4: Broadcast message formats.

2. Architecture Description

According to the MIL-STD-1553B the functional architecture of the core is shown in figure 5. The following architecture can be used as bus controller, remote terminal or bus monitor.

The total design is sub divided in to several blocks. The CPU writes all the required commands and data into the RAM for each transmission. There by reducing the load and wastage of time of CPU as the clocks are different. The CPU gets interrupts from the interrupt handler, which takes care of informing the CPU regarding the completion of transmission, errors during transmission and status of the terminal. For a message retransfer command it simply gives commands to RAM instead CPU, thereby reducing load on CPU. The interrupt handler Communicates with engine and takes signals, process them and produces required interrupts or commands.

The engine takes the bits from the registers block and produces enable, load, transmission and signals that are required for the remaining blocks. The engine gives the selection bits to the multiplexer depending on the bus activity. It gives the load signals to the register blocks. The engine also sets the mode of the core i.e. bus controller or remote terminal or bus monitor.

The host interface acts as Communication Bridge between the internal block and RAM; it passes the data on the RAM on to the internal data bus which is 16 bit wide. The internal data bus connects to internal registers and memory block. The registers block contains the control register, status register, error registers. Each register is 16 bit. The control register hold the control bits, the status register holds different status bits regarding transmission, memory block etc. The error contains the information of errors occurred.

The Tx memory block contains command registers which stores commands

.

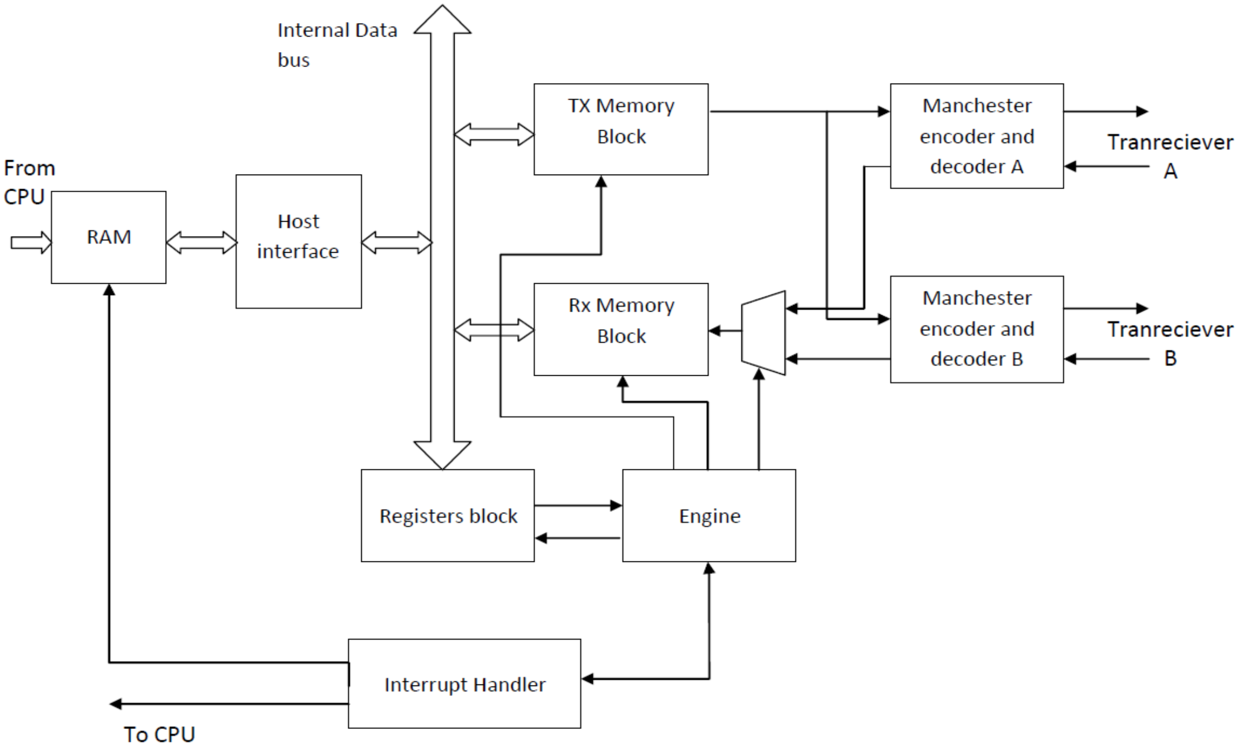


Fig 5: Architecture of the MIL-STD-1553B core.

to be transmitted during bus controller mode, status word register which stores the status words in remote terminal mode and FIFO that stores data to be transmitted. The Rx memory block contains the status word registers that stores the status words received for bus controller mode, command word registers for remote terminal mode and FIFO to store the data received on the buses.

The Manchester encoder and decoder encode the data and commands that are to be transmitted in to bi phase Manchester II format. The Manchester II format provides a self clocking waveform in which the bit sequence is independent. Similarly it decodes the messages received in bi phase Manchester II format. There are two Manchester encoder and decoders each for bus A and bus B. The data is received from the transreceivers that are connected to thebuses.

3. Implementation Results

The IP core of MIL-STD-1553 has been written using the verilog HDL and implemented in Xylinx vertex-2,Spartan-3 FPGA and it has used area of 891(4- LUT count).The output on the bus for different types of transmissions is shown in figure 6.

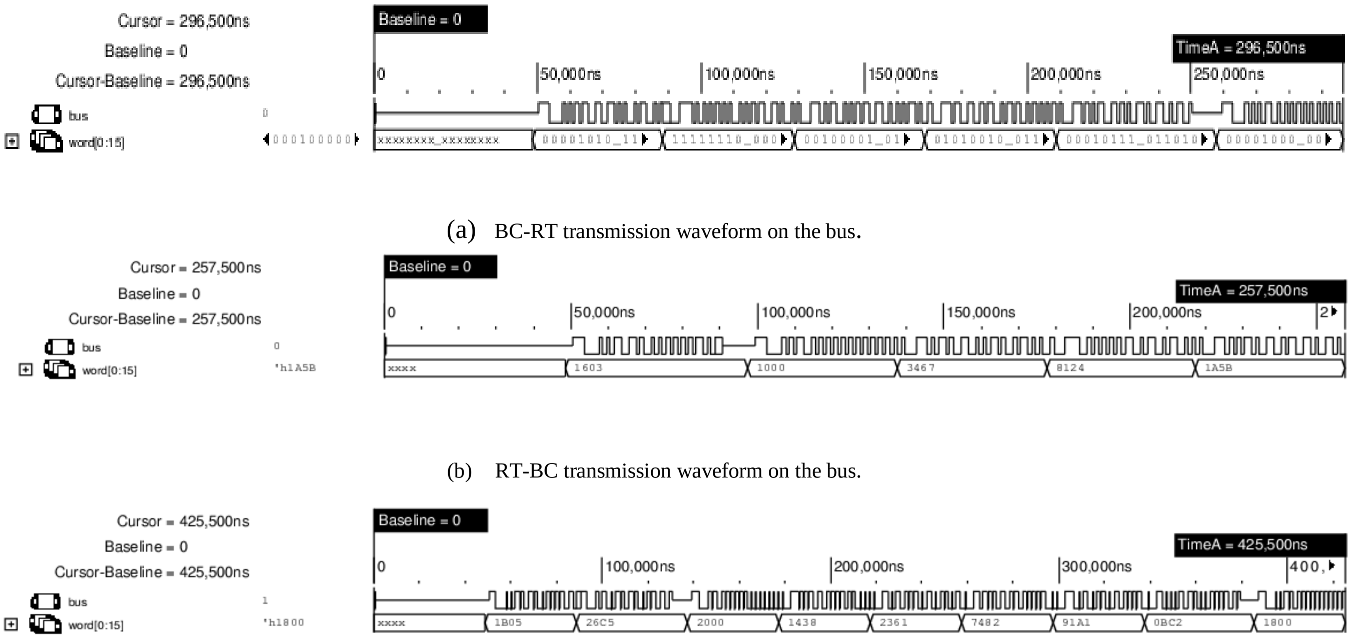
References

[1] MIL-STD-1553B NOTICE II. Department of Defence, washington D.C. 20360. 1986, 9.

[2] Design and Implementation of 1553B Bus Interface Board Based on PCI Bus ZHANG Rong-feng, ZHU Jian, XIA Wen-yuan, SUN Sheng-li 2008.

[3]. Design of the MIL-STD-1553B RT Based on BU-61580 WANG Haotong, JIA Hui-fu,GAO Rui-qian .Journal of Naval Aeronautical and Astronautical University. 2008.23 (4):439-441.

[4]. MIL-STD-1553A/B Designer's Guide, DDC Data Device Corporation



(a) BC-RT transmission waveform on the bus.

(b) RT-BC transmission waveform on the bus.

(c) RT-RT transmission waveform on the bus.

Figure 6: The waveforms on bus for different type's transmission, in Manchester II format.

Multimedia-based Medicinal Plants Sustainability Management System

Zacchaeus Omogbadegun[1§], Charles Uwadia[2], Charles Ayo[3], Victor Mbarika[4], Nicholas Omoregbe[5] , Efe Otofia[6] , Frank Chieze[7]

[1]Computer and Information Sciences Department, College of Science and Technology, Covenant University,
Ota, Ogun State, Nigeria

[2]Department of Computer Sciences, University of Lagos,
Lagos, Nigeria

[3]Computer and Information Sciences Department, College of Science and Technology, Covenant University,
Ota, Ogun State, Nigeria

[4]Southern University and A&M College,
Baton Rouge, LA 70813, USA

[5]Computer and Information Sciences Department, College of Science and Technology, Covenant University,
Ota, Ogun State, Nigeria

[6]Computer and Information Sciences Department, College of Science and Technology, Covenant University,
Ota, Ogun State, Nigeria

[7]Electrical and Information Engineering Department, College of Science and Technology, Covenant University,
Ota, Ogun State, Nigeria

[§]Corresponding author

Abstract

Medicinal plants are increasingly recognized worldwide as an alternative source of efficacious and inexpensive medications to synthetic chemo-therapeutic compound. Rapid declining wild stocks of medicinal plants accompanied by adulteration and species substitutions reduce their efficacy, quality and safety. Consequently, the low accessibility to and non-affordability of orthodox medicine costs by rural dwellers to be healthy and economically productive further threaten their life expectancy. Finding comprehensive information on medicinal plants of conservation concern at a global level has been difficult. This has created a gap between computing technologies' promises and expectations in the healing process under complementary and alternative medicine. This paper presents the design and implementation of a Multimedia-based Medicinal Plants Sustainability Management System addressing these concerns. Medicinal plants' details for designing the system were collected through semi-structured interviews and databases. Unified Modelling Language, Microsoft-Visual-Studio.Net, C#3.0, Microsoft-Jet-Engine4.0, MySQL, Loquendo Multilingual Text-to-Speech Software, YouTube, and VLC Media Player were used.

***Keywords**: Complementary and Alternative Medicine, conservation, extinction, medicinal plant, multimedia, phytoconstituents, rural dwellers*

1. Introduction

Plants and animals hold medicinal, agricultural, ecological, commercial and aesthetic/recreational value. Some plants of medicinal value: *Anacardium occidentale (*Cashew nut), *Azadirachta indica (*Neem), *Allium sativum* L.(Garlic), and *Zingiber officinale* Roscoe (Common Ginger), are shown in **Fig. 1**.

Neem
Fruits of Azadirachta indica
Azadirachta indica tree
Zingiber officinale rhizome
Zingiber officinale plant
Qinghao

Fig. 1 Some plants of medicinal value

Medicinal plants have become the most important source of life-saving drugs for the majority of the world's population. Medicinal plants harvested from the wild remain of immense importance for the well-being of millions of people around the world. Over 70,000 plant species are thought to be medicinal **[1].**

Medicinal plants are considered a source of various alkaloids and other chemical substances essential for mankind. Over 80% of the US public uses nonconventional practices and complementary medicines adjunctive to conventional medical care. According to the World Health Organization, over 80% of the people in developing countries depend upon traditional medicine for their primary health care **[2].**

Africa has been and continues to be a significant source of medicinal and aromatic plants and botanicals to the world's food, drug, herb and dietary supplement market. About 50% of drugs used in modern medicine are of plant origin. About 80% of Africa's population rely on medicinal plants for their health needs confirming that medicinal plant preparations have been identified as alternative remedies for several diseases **[3].** The active principles of many plant species are isolated for direct use as drugs, lead compounds or pharmacological agents **[4].**

The medicines for internal use prepared in the traditional manner involve simple methods such as hot- or cold-water extraction, extraction of juice after crushing, powdering of dried material, formulation of powder into pastes via such a vehicle as water, oil or honey, and even fermentation after adding sugar source. The range of products that could be obtained from medicinal plants is given in **Fig**.**2 [5].**

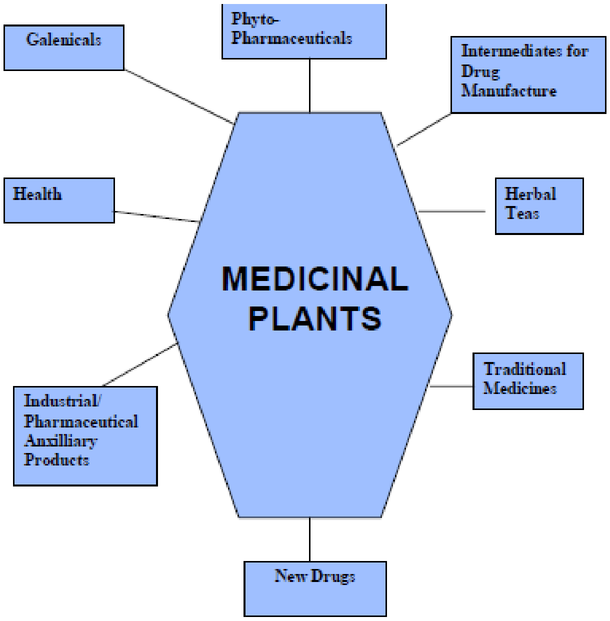


Fig. 2 Industrial Uses of Medicinal Plants [5]

Low accessibility to and affordability of orthodox medicine by rural dwellers and their need to keep healthy to be economically productive have led to their dependence on medicinal plants to remedy afflictions **[6].** In Nigeria herbal practices, the practitioners claim that plant parts possess various phytochemicals which exhibit diverse pharmacological and biological responses and diversities. Nigeria is a country stepped in the use of and belief in traditional medicines in which plants play a major role **[7].**

1.1 Statement of the Problem

Emerging new infectious, chronic and drug-resistant diseases have prompted scientists to look towards medicinal plants as agents for treatment and prevention.

Medicinal plants' species are threatened by habitat loss, climate change, and species-specific, multipurpose over-harvesting and logging leading to potential extinction of useful medicinal plants in the continent.

Securing supplies of quality products before the over-harvesting of wild stocks depletes the resource constitutes a concern.

Declining wild stocks of medicinal plants are accompanied by adulteration and species substitutions, which in turn reduce efficacy, quality and safety.

Imperceptibly, these medicinal plants' sustainability remains in jeopardy creating a gap between computing technologies' promises and expectations in the healing process under complementary and alternative medicine (CAM).

Difficulty encountered in finding comprehensive information on medicinal plants at a global level to promote scientific research towards obtaining clues and discovery of potential lead compounds and novel therapeutics.

Improvement in the quality of life of the rural poor; development of traditional medicines and reduction of the overexploitation of plants are inevitably desirable.

1.2 Research Questions

The following Research Questions have been raised for attention in this paper: (1) How can empirical knowledge of medicinal plant uses, often held by an older generation of healers in remote areas, be accumulated, stored and transmitted to next generations without compromising their intellectual property rights? (2) How can consumers be protected from false information or the use of products with negative side-effects? (3) How can sustainable wild sourcing be implemented – or the medicinal plants be 'domesticated' – to secure supplies of quality products before the over-harvesting of wild stocks depletes the resource?

1.3 Objectives of the Research

Our objectives included (1) providing a platform for a multidisciplinary team of scholars and healthcare services providers / CAM practitioners for information exchange in African healing process seamlessly;
(2) designing a framework for conserving, protecting and propagating medicinal plants, animals and cultural sites across the African continent; and
(3) implementing a system which would facilitate efficient knowledge discovery on medicinal plants with voice/video features on a multimedia platform.

2. Patients' Healthcare Requirements

Patients' requirements for healthcare include treatment and care that work, good relationship with practitioner, provision of information, and remaining in control of treatment. Complementary and alternative medicine continues to attract patronage due to patients' dissatisfaction with conventional health care, a desire for greater control over one's health, and a desire for cultural and philosophical congruence with personal beliefs about health and illness **[8].**

Nigeria is rich in biodiversity. The country is endowed with a variety of plant and animal species. As reflected in **Table 1,** there are about 7, 895 plant species identified in 338 families and 2, 215 genera **[9].**

Table 1 Plant species in Nigeria [9]

Groups Of Plants	Families	Genera	Species
Algae	67	281	1335
Lichens	-	14	17
Fungi (Mushrooms)	26	60`	134
Mosses	-	13	16
Liverworts	-	16	6
Pteridophytes	27	64	165
Gymnosperms	2	3	5
Chlamydosperms	2	2	6
Monocotyledons	42	376	1575
Dicotyledons	172	1396	4636
Total	**338**	**2215**	**7895**

The expanding trade in Medicinal Plants has serious implications on the survival of several plant species, with many under serious threat to become extinct. Recently however, attention is turning back to natural products as drug sources, since they have been so successful in the past. Modern medicine depends on biological materials as an incomparable source of molecular diversity. Against this backdrop, almost half the world's plant species may be threatened with extinction; cures as yet undiscovered may exist in plants as yet un-described - and which may never be described. Promising drug sources are also found in the sea - sponges, sea squirts and algae for example, are all sources of drugs undergoing clinical studies. Plants are the structural anchors of the ecosystems in which these organisms live. The rapid loss of plant life has far-reaching consequences, and their loss will adversely affect future drug discovery **[1].**

2.1 Medicinal Plants Endangered

Due to irresponsible human acts of mass destruction of forests worldwide, we are losing flora at an alarming rate. Unless we act immediately to preserve the medicinal plants, plants with nutraceutical values, future generations will loose tremendous health and wellness benefits from nutraceutical herbs that we are enjoying now **[10].**

Despite the long tradition of usage of medicinal plants, their proven efficacy, and lack of affordable alternatives, the continued availability of many of these plants is in jeopardy as their species are threatened by habitat loss, climate change, multipurpose over-harvesting, and logging.

Many medicinal plants are being destroyed at an unprecedented rate and are threatened with extinction. The destruction of plant species is occurring at a rate unmatched in geological history. Current extinction rates are at least 100 to 1,000 times higher than natural background rates, with a quarter of the world's coniferous trees in jeopardy, and as many as 15,000 medicinal plants threatened **[11]**.

In USA, approximately 250,000 species of flowering plants, it is estimated that some 60,000 of these may become extinct by the year 2050, and more than 19,000 species of plants are considered to be threatened or endangered from around the world. More than 2000 species of plants native to the United States are threatened or endangered, with as many as 700 species becoming extinct in the next 10 years **[12].**

In Pakistan, it was observed that 49 medicinal plants species belonging to 32 different families were sold in local markets and thus playing a role in uplifting the socioeconomic conditions of the area. It was observed that out of these 49 medicinal plants, 24 plant species are threatened (9 Endangered, 7 Vulnerable and 8 Rare) due to excessive collection from the wild. These plants are also used locally for curing different ailments. In most cases, the market availability status of these medicinal plants have increased, showing an increased inclination of local people towards medicinal plants collection and increased dependency of local population on medicinal plants trade. A brief set of information about these plants is given in **Table 2 [13].**

Table 2 Folk medicinal uses, market availability status, conservation status of some important medicinal plants of Swat, Pakistan [13] (Extracts)

Plant Material	Family	Part Used	M S	C S
Acorus calamus L.	Araceae	Whole plant	P	E
Berberis vulgaris Linn	Berberidaceae	Whole plant	P	E
Dioscorea deltoidea Wall.	Dioscoreaceae	Tubers	D	E
Polygonatum verticillatum All.	Liliaceae	Rhizome	P	E
Paeoniaceae *Paeonia emodi* Wall. ex Hk.f.	Paeoniaceae	Rhizome/ seeds	P	E
Podophyllum hexandrum Royle	Podophyllaceae	Rhizome	P	E
Bistorta amplexicaulis (D.Don) Greene	Polygonaceae	Rhizome	P	E
Bergenia ciliate (Haw) Sternb.	Saxifragaceae	Leaves Rhizome	I	E
Valeriana jatamansi Jones	Valerianaceae	Rhizome	D	E
Adiantum capillus-veneris L.	Adiantaceae	Fronds	I	V
Pistacia integerrima Stew.ex Brand	Anacardiaceae	Leaves	I	V
Berberis lyceum Royle	Berberidaceae	Whole plant	I	V
Ephedra gerardiana Wall. ex Stapf	Ephedraceae	Fruit/ leaves	I	V
Colchicum luteum Baker.	Liliaceae	Rhizome/ seeds	I	V

Legend: MS = Market Status [D = Decreased, I = Increased, P = Persistent], CS = Conservation Status [E = Endangered, V = Vulnerable]

The downturn in the Nigerian economy and inflationary trend has led to the excessive harvesting of non-timber forest products for various uses. Some of these species are now threatened. Examples as reflected in **Table 3** are H*ymenocardia acida, Kigelia africana, and Cassia nigricans* **[9].**

Table 3 Threatened Biodiversity Species in Nigeria [9]

SPECIES	MAIN USES	STATUS
PLANTS		
Milicea excelsia	Timber	Endangered
Diospyros elliotii	Carving	Endangered
Triplochiduiton scleroxylon	Timber	Endangered
Mansoiea altissinia	Timber	Endangered
Masilania accuminata	Chewing stick	Endangered
Garcina manni	Chewing stick	Endangered
Oucunbaca aubrevillei	Trado-medical	Almost Extinct
Erythrina senegalensis	Medicine	Endangered
Cassia nigricans	Medicine	Endangered
Nigella sativa	Medicine	Endangered
Hymenocardia acida	General	Endangered
Kigelia africana	General	Endangered

The Red Data Book of India has 427 entries of endangered species of which 28 are considered extinct, 124 endangered, 81 vulnerable, 100 rare and 34 insufficiently known species. In West Africa, Nigeria has the highest number of threatened species (119), followed by Ghana (115), Côte d'Ivore (101), Liberia (46) and Sierra Leone (43) **[14].**

A total of 54 different tree species (24 families) were identified in Ala, 41 species (21 families) in Omo and 55 species (20 families) in Shasha Forest Reserves of Nigeria. The most prevalent species in the ecosystem was *Strombosia pustulata*, while the family Leguminosae had the highest number of species. 84% of the species are regarded as rare or threatened with extinction while 16% were relatively abundant. Wanton removal of plant and animal species through over-harvesting activities is very inimical to biological conservation and has led to loss of biodiversity and extinction especially of many natural species with narrow range **[15]**. The status of medicinal plants under regular trade in the rainforests is presented in **Table 4**.

Table 4 Respondent Opinions on the Status of Traded Medicinal Plants in Community Forests [15] (Extracts)

Medicinal plants	Opinion of respondents (%)				
	Engd	**Thrtd**	**Rar**	**Comn**	**%**
Alchornea cordifolia	44	24	20	12	100
Ananthus montanus	32	54	10	4	100
Bridelia ferruginea	30	48	22	24	100
Callichilia barteri	42	22	20	16	100
Canarium schweinfurthii	32	28	24	16	100
Cissus aralioides	34	48	10	8	100
Cocholospermum planchonni	44	26	16	14	100

Combretum smeathmanii	48	23	24	2	100
Enantia chloratha	44	30	14	12	100
Ocimum gratissimum	46	24	20	10	100
Rauvolfia vomitoria	24	64	8	4	100
Rauvolfia vomitoria	22	60	14	4	100
Rothmannia hispida	46	38	10	6	100
Sanseuieria guineense	24	60	14	2	100
Struchium spargonophora	32	48	22	18	100
Thorningia sanguinea	24	42	20	14	100
Uraria picta	44	22	14	20	100
Zingiber officinale	56	24	10	10	100

Legend: Engd = Endangered, Thrtd = Threatened, Rar = Rare, Comn = Common

The data were based on the opinions of stakeholders (respondents) directly involved in harvesting medicinal plants for the market. The data in **Table 4** revealed that 45 medicinal plants were traded on regular basis in the rainforest of Nigeria. Out of these, 8 medicinal plants were endangered, 12 species were rare and 8 species were threatened, while 17 were common or abundant in natural forests **[16].**

Plants and animals are responsible for a variety of useful medications. In fact, about forty percent of all prescriptions written today are composed from the natural compounds of different species. These species not only save lives, but they contribute to a prospering pharmaceutical industry worth over $40 billion annually. Unfortunately, only 5% of known plant species have been screened for their medicinal values, although we continue to lose up to 100 species daily. The Pacific yew, a slow-growing tree found in the ancient forests of the Pacific Northwest, was historically considered a "trash" tree (it was burned after clearcutting). However, a substance in its bark taxol was recently identified as one of the most promising treatments for ovarian and breast cancer. Additionally, more than 3 million American heart disease sufferers would perish within 72 hours of a heart attack without digitalis, a drug derived from the purple foxglove **[17].**

All over the world, much importance is being given to the documentation of knowledge on traditional health remedies, conservation and sustainable use of biodiversity, cultivation, value addition, and development of standards for indigenous drugs. Unfortunately it is difficult to find comprehensive information in this sector at a global level **[14].** Hawkins **[1]** provided sources of medicinal plants of conservation concern worldwide.

2.2 Related Works

Efforts have been and still being made by various research institutions, botanic gardens in published literature and databases (including **[1], [13], [18] – [34]),** and herbarium to document and conserve medicinal plants.

The New York Botanical Garden currently grows ten species of plants on the Federal Endangered Species List. They are striving to preserve rare and endangered plants and participate with other institutions in doing this. The National Collection of Endangered Plants contains seeds, cuttings, and whole plants of 496 rare plant species native to the United States. The collection is stored at 25 gardens and arboreta that form part of the Center for Plant Conservation (CPC). The Royal Botanic Gardens at Kew, United Kingdom, support six *ex situ* and *in situ* conservation projects. The activities range from acting as the U.K. scientific Authority for Plants for Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES), cooperating in the recovery and reintroduction of endangered species, and aiding in the production of management plans for sustainable development and protected areas. The Wrigley Memorial and Botanical Gardens at Catalina Island, California, is still another example. The Gardens' emphasis is on California island endemic plants. Many of these plants are extremely rare, with some listed on the Endangered Species List **[12]**.

MEDPHYT **[35]** was built to collect data on the complete European pharmaceutical and toxicological plant world whose representatives were determined by medical and therapeutic benefit. The focus of the database content was the plant with description of their botanical characteristics, and history of discovery of therapeutic use, etymology, and synonyms. Besides the botanical characterisation there was information on both medical relevant biochemical compounds and their physicochemical characteristics, and toxicological as well as pharmaceutical facts. Its future outlook included comprehensive 3D visualisation techniques, implementation of the system for mobile systems and expansion of the content. This would require some additional work on both new scientific approaches to the action of mode of the drugs on molecular level and the database system MEDPHYT. Addition of further tables for animals, bacteria and fungi would also be required.

None of these attempts incorporated text-to-speech, voice/video, and multilingual features to address different cultures, languages, and audiences comprehensively on a multimedia platform. This gap limits knowledge discovery on useful medicinal plants from different areas of origin to be propagated elsewhere in addressing common diseases and ailments. Promotion of scientific research towards obtaining clues and discovery of potential lead compounds and novel therapeutics from medicinal plants has consequently been retarded.

3. Materials and Methods

3.1 Data Collection

The criteria used for our data collection were centered on multipurpose uses in primary health care, needs for cultivation packages, potential in processing at the primary level, knowledge of availability and threat to the wild resources, and the need for collaborative work for understanding the patterns of diversity in relation to medicinal value and use.

The ethnobotanical data on plant species, families, vernacular names, prevalence, sustainable availability, plant parts used, medicinal usage by local communities and modalities of use, conditions for cultivation, phytochemical and pharmacological properties were collected, recorded, and discussed.

Medicinal plant materials were obtained randomly through personal contacts in the field, forestry-and-plant-science-based research institutions, local markets (*elewe-omo*), and at the homes of complementary and alternative medical practitioners (*babalawos*) in January - March 2010, June – September 2010, and December 2010 - February 2011 in Ota and Abeokuta towns in Ogun State of Nigeria.

Medicinal plant uses were discussed in detail with informants, after seeking prior informed consent from each respondent. Two hundred and fifty complementary and alternative medicine practitioners who know and use medicinal plants for treating various diseases were interviewed.

Following a semi-structured interview technique randomly administered, respondents were asked to provide detailed information about the vernacular plant name in Yoruba language; plant properties; harvesting region; ailments for which a plant was used; best harvesting time and season; plant parts used, as well as mode of preparation and application.

Older individuals, local medicine men or herbalists and others who claim to have effective prescriptions were interviewed. All interviews were carried out in Yoruba language, with at least one of the authors present. Three of the authors were fluent in Yoruba language, and no interpreter was needed to conduct the interviews.

Table 5 shows some of the medicinal plants' characteristics gathered, while **Table 6** presents samples of the collected medicinal plants.

Table 5 Data Collection Parameters (Extracts)

Scientific/BotanicalName	***Family Name***	**CommonName**	**Synonyms**	***Local Names(Yoruba Lang)***	**Description**	**Medicinal Uses**	***Parts Used***	**Area(s) of Origin**	***Preparations / Dosage***	***Contraindications***	***Phytoconstituents***	**Adverse Reactions**	**Toxicity**	**Pharmacology**	***Drug interactions***	***Picture***	***Published Source(s)***

*Extinct, Vulnerable, Threatened, Endangered, Available, Rare

Table 6 Extracts from the collected/analysed medicinal plants

Name			Parts Used	Use	Photo
Species	Family	Yoruba Lang			
ACA	*facx*	*Jinwini*	*root decoction*	*WI*	
AGR	*fagx*	*Imi-esu, Akayunyun*	*leaf decoction*	*URT* *WI*	
ALL	falx	Alubosa ayu	Root	STR EYE	
ASP		*Aluki, Eye-kosun-Dangi*	*Root*	*MI*	
ELY	*fact*	*Ewe-Eso*	*Whole plant.*	*GNO* *IMP* *INF*	
EUP	*feux*	*Orowere, Enuopire Enukopure*	*Leaves, exudate*	*DMT* *INF*	
FIC	*fmox*	*Opoto, Farin bauree, Anwerenwa*	*Leaves stem, root, fruits*	OED LEP EPL RIC INF	
ZNG	fznx	Jinja, Atale, Atalekopa	Rhizome Root	AST PIL, HEP OBE ANN CAN DYS	

Legend: ANA = Anaemia, AST = Asthma, CAN = Cancer,

DMT = Dermatosis, DYS = Dysmenorrhoea, EPL = Epilepsy, EYE = Eye pain, GNO = Gonorrhoea, HEP = Hepatitis, IMP = Impotence,
INF = Infertility, PIL = Piles, LEP = Leprosy, MI = Male Infertility,
OBE = Obesity, OED = Oedema, RIC = Rickets, STR = Stroke,
URT = Urinary Tract Infetcions, *WI* = Women Infertility
ACC = *Acalypha villicaulis Hoschst*, *AGR* = *Ageratum conyzoides L*, ALL = *Allium sativum* L., ASP = *Asparagus racemosa*, ELY = *Elytraria marginata*, EUP = *Euphorbia laterifolia*, FIC = *Ficus capensis* Thunb , ZNG = *Zingiber officinale* Rosc
facx = *Euphorbiaceae*, *fagx* = *Asteraceae*, falx = Alliaceae,
fact = *Acanthaceae*, *feux* = *Euphorbiaceae*, *fmox* = *Moraceae*,
fznx = Zingiberaceae

Additional data on medicinal plants were also collected from published literature: **[18] - [25];** phytochemical databases; and reputably cognate internet-based medicinal plants' sources including **[26]-[34]** and shown in **Fig. 3**.

Medicinal_Plants_in_Nigeria
www.medicinalplantsinnigeria.com,
(Odugbemi, 2010)

19 PhD Progress Report - OmogbadegunZO (2011) Tuesday, 16 August 2011

Fig. 3 Medicinal Plants in Nigeria [30]

3.2 Software Development

Guided by Threatened Species Categorization Standards in **Fig. 4 [1]**, adapting and modifying MEDPHYT of **Fig. 5 [35]** and DeeprootPlantBase of **Fig. 6** prototypes**,** the following were used: Visual Studio.Net and C# 3.0 Programming Language for creating the application, Microsoft Access for creating and querying the Database, HTML for displaying the plant Information in a text format, and Microsoft Jet Engine 4.0 to connect the application to the Database. Loquendo Multilingual Text-to-Speech Software was incorporated for audio content of the plant's salient characteristics, while YouTube and VLC Media Player were used for showing and playing the video of each plant chosen.

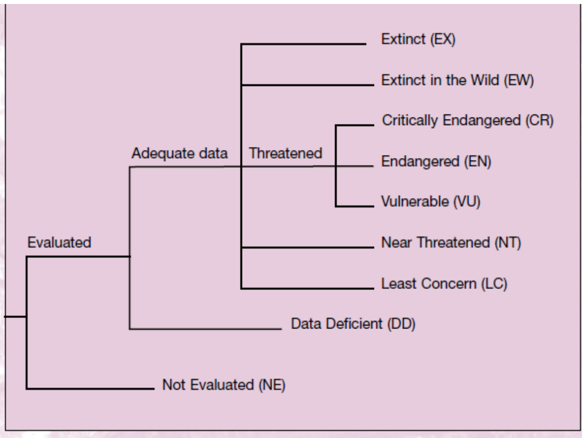
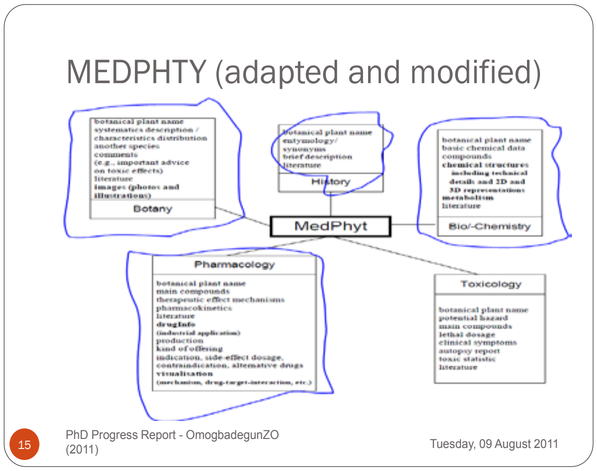
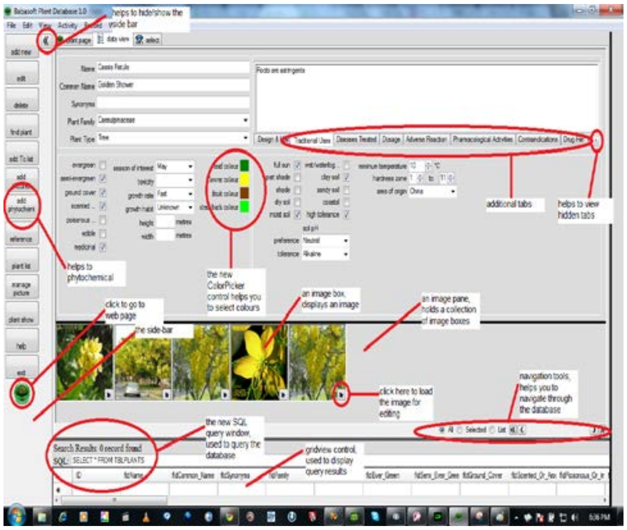


Fig. 4 The 2007 IUCN Red List of threatened species categories [1]


MEDPHTY (adapted and modified)

15 PhD Progress Report - OmogbadegunZO (2011) Tuesday, 09 August 2011

Fig. 5 MEDPHTY adapted [35]


Name: Cassia Fistula
Common Name: Golden Shower
Plant Family: Caesalpinaceae
Plant Type: Tree
Roots are astringents
Search Results: 0 record found
SQL: SELECT * FROM TBLPLANTS

Babasoft Plant Database 1.0
File Edit View Activity
helps to hide/show the side bar
add new
edit
delete
find plant
add To list
add phytochem
reference
plant list
manage picture
plant show
help
exit
data view
select
Design & Use
Traditional Uses
Diseases Treated
Dosage
Adverse Reaction
Pharmacological Activities
Contraindications
Drug Her
additional tabs
helps to view hidden tabs
helps to phytochemical
the new ColorPicker control helps you to select colours
an image box, displays an image
an image pane, holds a collection of image boxes
click to go to web page
the side-bar
click here to load the image for editing
navigation tools, helps you to navigate through the database
All
Selected
List
the new SQL query window, used to query the database
gridview control, used to display query results
ID
fldName
fldCommon_Name
fldSynonyms
fldFamily
fldEver_Green
fldSemi_Ever_Green
fldGround_Cover
fldScented_Or_Not
fldPoisonous_Or_It

Fig. 6 Enhanced Medicinal Plants Database after adaptation from Deeproot Plantbase

4. Results

Over 250 medicinal plants were collected, analysed, and discussed. The entire plants with their flowers, fruits, and roots were collected and photo/snap taken. All the collected plants' details were used to populate the database of the developed system to address the medicinal plants' extinction problem as reflected in **Figs 6 - 7.**

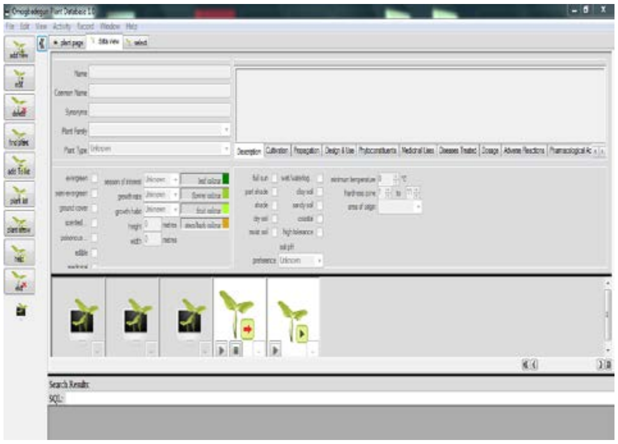
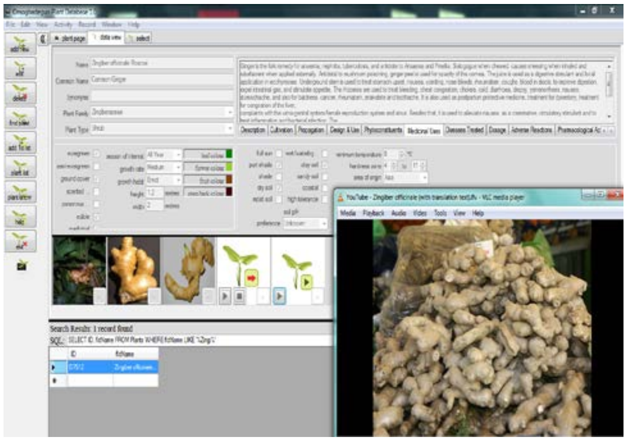


Fig. 7 Multimedia-based Medicinal Plants Sustainability Management System (ONI_MMPSMS)

Details of each plant included scientific name, common name(s), family, multiple areas of origin, phytoconstituents, traditional medicinal uses, diseases treated, pharmacological activities, prescriptions, dosage, mode of preparation and administration, adverse reactions, toxicity, contraindications, drug-herb interactions, graphics, text-to-speech of contents, images, phytochemical structure, etc. A typical example extracted from **Table 6** and **Fig. 8**, *Zingiber officinale* Rosc (Common Ginger, - *Jinja, Atale, Atalekopa),* reflects Ginger (Zingiber Officinale, *Zingiberaceae* family) as originating from southern Asia. Nowadays, it is cultivated and commercialized around the world, particularly in China, India, Indonesia and Africa to demonstrate multiple areas of availability.



Search Results: 1 record found

SQL: SELECT ID, fctName FROM Plants WHERE fctName LIKE '%Zing%'

Fig. 8 Enquiry from ONI_MMPSMS showing Diseases treated with a plant

The three major producing countries are India, China, and Nigeria. Technological and socioeconomic factors for cultivating *Zingiber officinale* Rosc were captured and stored. Among others, it was evident that more than one medicinal plant could be used in treating the same disease/ailment. For example, in **Table 6**, *Acalypha villicaulis Hoschst* and *Ageratum conyzoides* which differ only in the parts used treat women infertility. Similarly, *Elytraria marginata* and *Euphorbia laterifolia* handle infertility in both male and female.

In a similar vein, a single medicinal plant has multi-purpose use in handling more than one disease/ailment. Examples from **Table 6** showed *Elytraria marginata (for* Gonorrhoea, impotence, infertility), *Euphorbia laterifolia (for* Dermatosis, infertility*), Ficus capensis* Thunb *(for* Dysentery, oedema, leprosy, epilepsy, rickets, infertility), *Zingiber officinale* Rosc (for asthma, stimulant, piles, hepatitis, liver diseases, obesity, typhoid, anaemia, cancer, dysentery, Dysmenorrhoea).

Comprehensively on medicinal uses as documented in the database, Ginger is the folk remedy for anaemia, nephritis, tuberculosis, and antidote to Arisaema and Pinellia. Sialogogue when chewed, causes sneezing when inhaled and rubefacient when applied externally. Antidotal to mushroom poisoning, ginger peel is used for opacity of the cornea. The juice is used as a digestive stimulant and local application in ecchymoses. Underground stem is used to treat stomach upset, nausea, vomiting, nose bleeds, rheumatism, coughs, blood in stools, to improve digestion, expel intestinal gas, and stimulate appetite. The rhizomes are used to treat bleeding, chest congestion, cholera, cold, diarrhoea, dropsy, dysmenorrhoea, nausea, stomachache, and also for baldness, cancer, rheumatism, snakebite and toothache. It is also used as postpartum protective medicine, treatment for dysentery, treatment for congestion of the liver, complaints with the urino-genital system/female reproduction system and sinus. Besides that, it is used to alleviate nausea, as a carminative, circulatory stimulant and to treat inflammation and bacterial infection. The Commision E approved the internal use of ginger for dyspepsia and prevention of motion sickness. The British Herbal Compendium indicates ginger for atonic dyspepsia, colic, vomiting of pregnancy, anorexia, bronchitis and rheumatic complaints. European Scientific Cooperative on Phytotherapy (ESCOP) indicates its use for prophylaxis of the nausea and vomiting of motion sickness and to alleviate nausea after minor surgical procedures.

From this work, on **contraindications** for *Zingiber officinale* Rosc, it has been documented that users should consult physician before using ginger preparations in patients with blood coagulation disorders, taking anticoagulant drugs or with gallstones.

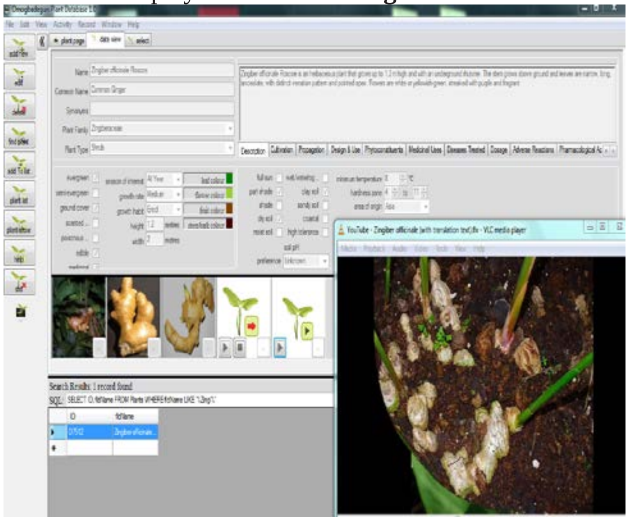
With respect to **toxicity**, ginger is a safe drug without any adverse reactions and has a wide range of utility. However, dried rhizomes during pregnancy should be avoided.

Fresh and dry ginger are tolerant and could be used as such. Generally, ginger is not subjected to any **purification** methods. Methods of purification for dry ginger and fresh juice are available from *Arogyakalpadruma* (an Ayurvedic text that concentrates on pediatrics). Purification of ginger may therefore be intended only for pediatric use, that is, to reduce the potency and pungency for infant use **[36].**

With respect to **drug-herb interactions**, it has documented that *Zingiber officinale* Rosc interacts with anticoagulants such as heparin, warfarin, drugs used in chemotherapy and ticlopidine. Ginger taken prior to 8-MOP (treatment for patients undergoing photopheresis) may substantially reduce nausea caused by 8-MOP. Ginger appears to increase the risk of bleeding in patients taking warfarin. However, ginger at recommended doses does not significantly affect clotting status, the pharmacokinetics or pharmacodynamics of warfarin in healthy subjects. Ginger also significantly decreased the oral bioavailability of cyclosporine. All the above characteristics have been captured for each medicinal plant as exemplified in **Fig. 8.**

The novelty in this work included multilingual text-to-speech (voice) and video features in a collaborative virtual environment for seamless exchange of information among scholars and practitioners of complementary and alternative medicine.

In addition, users of the system could search for any plant by any combination of its characteristics. The matched items are displayed in a separate window for the user to pick the desired plant. Furthermore, an interface exists for a sophisticated computer database user to use SQL SELECT statement in his/her search exercise. Plant search by other parameters such as area of origin is featured. Standard update actions in software engineering are provided. The video of the salient details of the retrieved plant could be viewed, and the audio feature could also be played as shown in **Figs 8 - 9**.



Name: Zingiber officinale Roscoe
Common Name: Common Ginger
Plant Family: Zingiberaceae
Plant Type: Shrub

Search Results: 1 record found

Fig. 9 Screen shot of ONI_MMPSMS features

Specific database of medicinal plants for each disease could be extracted and expanded from the large system.

5. Discussion

Conservation of threatened medicinal plants has become an increasingly important role. Medicinal plants' value that complements orthodox medicine in affordable manner has been documented for sustainability. The information made available on multiple areas of origin for cultivation elsewhere in solving medicinal plants' extinction challenge would also promote scientific research towards obtaining clues and discovery of potential lead compounds and novel therapeutics.

Beneficiaries of the software would include computer scientists, information technologists, biochemists, biologists, chemists, ethnobotanists, phytochemists, phytopharmacists, physicians and African traditional doctors (typically herbalists) to exchange information in African healing process. New plants would be added seamlessly as characteristics are assembled.

Among other characteristics, the following medicinal uses of *Zingiber officinale* Rosc published by **[37]** have been documented in the database as shown in **Fig. 8**: "Ginger (rhizome of *Zingiber officinale*) is a well known herb for its culinary and wide range of medicinal uses and is considered an essential component of the kitchen pharmacy. More commonly, ginger has been traditionally used in disorders of the gastrointestinal tract, as a stomachic, laxative, sialogogue, gastric emptying enhancer, appetizer, antiemetic, antidyspepsic, antispasmodic, and antiulcer agent with sufficient scientific support. Similarly, ginger has been shown to exhibit anti-inflammatory, hypoglycemic, antimigraine, antioxidant, hepatoprotective, diuretic, hypocholesterolemic, and antihypertensive activities.

Recently, ginger has gained wide attention for its therapeutic role as a safe and effective preventive treatment option for nausea and vomiting of pregnancy. Ginger has a long history of safety, as it has been used for centuries not only for medicinal purposes, but also as a food and spice. Although some health risk and safety concerns exist in the literature about its use by pregnant women, the clinical evidence of harm is lacking. Ginger might, therefore, be used as an effective treatment option for nausea and vomiting during pregnancy" **[37].** It could be deduced that *Zingiber officinale* Rosc could serve as a single medicine for internal use, as an ingredient in compound medicines, for external use, as an adjuvant, as an antidote, and for the purification of some mineral drugs.

6. Conclusions

The emergence of new infectious, chronic and drug-resistant diseases have prompted scientists to look towards medicinal plants as agents for treatment and prevention to foster high-quality and high-efficiency primary care. A system that documents and maintains comprehensive details on medicinal plants has been discussed. This work has provided the general public valuable insights into availability of and the traditional use of plants as medicines. This research effort would no doubt lead to a greater sense of confidence in many of the leading botanical raw materials of African origin in the medicinal plant trade. With comprehensive information provided on medicinal plants, this system would serve as a vehicle for critical gap-filling in research, study and application of research results in medicinal plants. Finally, the work provides a database for full scientific research towards obtaining clues and discovery of novel therapeutics.

6.1 Recommendations

In view of the multi-purpose use of medicinal plants in enhancing health, the following strategies to bridge the gap in the sustainability of ginger advanced by **[37]** are highly recommended for all medicinal plants:

- Enhancement of potential and realizable productivity through an integrated system of cultivation using high-yielding and resistant varieties, plant nutrient management, production technology suited to different agroecological situations and cropping systems, and need-based plant protection measures are future areas of increasing relevance in boosting ginger production.
- Resistance breeding against devastating diseases such as soft rot and bacterial wilt incorporating genes from wild relatives using biotechnological tools or through the exploitation of somaclonal variations.
- Enhancement of quality and evolving and popularizing very efficient postharvest handling techniques including product diversification.
- Tailoring production to meet export needs and international requirements such as clean ginger.
- Organic ginger production is to be promoted as a large quantity of immature ginger and fresh ginger are being used for the production of fresh ginger products. Popularization of products such as ginger beer, ginger ale, ginger squash, and ginger tea among consumers can help to create demand for ginger, which in turn will boost the ginger economy.

6.2 Future Development

A mobile version of the system deployable on a 4G network technology is underway. This would be accessible by complementary and alternative medicine (CAM) practitioners in the rural and underserved population areas. The software would also be made available in Nigeria's principal local languages (Hausa, Igbo, and Yoruba) and French in no distant future.

Acknowledgment

Z.O. Omogbadegun, C.O. Uwadia, C.K. Ayo, V.W.Mbarika, N.Omoregbe, E.Otofia, and F. Chieze appreciate International Institute of Tropical Agriculture (IITA), Ibadan, Nigeria; Forestry Research Institute (FRI), Ibadan, Nigeria; Obafemi Awolowo University Herbarium, Ile-Ife, Nigeria; Dr. Abiodun Adebayo (Department of Biological Sciences, Covenant University, Ota, Nigeria), Geoff Looker (Deeproot Software Ltd, London, UK), Prof. Jiang Ding-Bi (Institute of Information on Traditional Chinese Medicine, China Academy of Chinese Medical Sciences, Beijing, P.R.China), and Prof. Yulan He (Informatics Research Centre, University of Reading, Reading RG6 6BX, UK).

References

[1] A. Hawkins, *Plants for life*: *Medicinal plant conservation and botanic gardens*, Richmond, U.K Botanic Gardens Conservation International, 2008

[2] B. Kasirajan; R. Maruthamuthu; V. Gopalakrishnan; K. Arumugam; H. Asirvatham; V. Murali; R. Mohandass; and A. Bhaskar, "A database for medicinal plants used in treatment of asthma", *Bioinformation* 2(3): 2007, pp. 105-106.

[3] F. Cho-Ngwa, M. Abongwa, M. Ngemenya, and K.D. Nyongbela. "Selective activity of extracts of *Margaritaria discoidea* and *Homalium africanum* on *Onchocerca ochengi", BMC Complementary and Alternative Medicine* 2010, 10:62doi:10.1186/1472-6882-10-62

[4] P.A. Babu; G. Suneetha; R. Boddepalli; V.V. Lakshmi; T.S. Rani; Y. RamBabu; and K. Srinivas. "A database of 389 medicinal plants for diabetes", *Bioinformation* 1(4):, 2006, pp. 130-131

[5] D.N.Tewari, Report of the Task Force on Conservation & Sustainable use of Medicinal Plants, Government of India, Planning Commission, March – 2000, http://planningcommission.nic.in/aboutus/taskforce/tsk_medi.pdf Accessed February 7, 2011.

[6] T. E. Mafimisebi and A. E. Oguntade. "Preparation and use of plant medicines for farmers' health in Southwest Nigeria: sociocultural, magico-religious and economic aspects", *Journal of Ethnobiology and Ethnomedicine*, 6:1, 2010.

[7] A. P. Ekanem and F. V. Udoh. The Diversity of Medicinal Plants in Nigeria: An Overview. In Chi-Tang Ho (Ed). *African Natural Plant Products: New Discoveries and Challenges In Chemistry and Quality (ACS Symposium Series),* Oxford University Press, USA, 2010

[8] S. B. Kayne, *Complementary and Alternative Medicine,* Second edition, London: Pharmaceutical Press, 2009.

[9] E. Oladipo, M. G. Ogbe, Norman Molta, David Ladipo, Gamaniel Shingu, et al. Current *Status* of Biodiversity in *Nigeria:* Nigeria First National Biodiversity Report, 2001. www.cbd.int/doc/world/ng/ng-nr-01-en.doc. Accessed March 12, 2011.

[10] S. Agrawal and A. Chakrabarti. Potential Nutraceutical Ingredients from Plant Origin. In Yashwant Pathak *(Ed). Handbook of Nutraceuticals Volume 1: Ingredients, Formulations, and Applications, CRC Press, Taylor* & Francis Group, *2010.*

[11] V. Brower, "Back to Nature: Extinction of Medicinal Plants Threatens Drug Discovery", *JNCI Journal of the National Cancer Institute*, 2008, 100 (12), pp. 838-839

[12] M.J. Bogenschutz-Godwin, J.A. Duke, M. McKenzie, and P.B. Kaufman. *Plant Conservation.* In L.J. Cseke, A. Kirakosyan, P.B. Kaufman, S.L. Warber, J. A. Duke and H. L. Brielmann (Eds). *Natural Products from Plants* Second Edition, Taylor & Francis Group, 2006, pp 503-534

[13] M. Hamayun, S.A. Khan, E.Y. Sohn, and In-Jung Lee. "Folk medicinal knowledge and conservation status of some economically valued medicinal plants of District Swat, Pakistan", L*yonia, a journal of ecology and application,* 2006, 11(2), pp 101-113

[14] K. Vasisht and V. Kumar. *Compendium of Medicinal and Aromatic Plants AFRICA*, United Nations Industrial Development Organization and the International Centre for Science and High Technology, 2004

[15] V. A. J. Adekunle. "Conservation of Tree Species Diversity in Tropical Rainforest Ecosystem of South-West Nigeria", *Journal of Tropical Forest Science,* 2006, 18(2): pp. 91–101

[16] G. J. Osemeobo. "Can the Rain Forests of Nigeria Sustain Trade in Medicinal Plants?", *International Journal of Social Forestry,* 2010, Volume 3, Number 1, pp. 66-80.

[17] www.endangeredspecie.com/Why_Save_.htm. Accessed January 25, 2011

[18] O.O. G Amusan, Herbal Medicine in Swaziland: An Overview. In Chi-Tang Ho (Ed). *African Natural Plant Products: New Discoveries and Challenges In Chemistry and Quality (American Chemical Society Symposium Series),* Oxford University Press, USA, 2010, pp. 26-44

[19] M.S. Dama, S.P. Akhand, and S. Rajender. "Nature versus nurture – plant resources in management of male infertility", *Frontiers in Bioscience* E2, 2010, 1001-1014, 1001, From Endocrinology Division, Central Drug Research Institute, Council of Scientific and Industrial Research, Lucknow, India - 226001

[20] T. L. Dog, Chaste Tree Extract in Women's Health: A Critical Review. In R. Cooper and F. Kronenberg (Eds). *Botanical Medicine: From Bench To Bedside,* Mary Ann Liebert, Inc 2010

[21] A.J. Hywood. Fertility Challenges. In A. J. Romm (Ed). *Botanical Medicine for Women.*, Churchill Livingstone, an imprint of Elsevier Inc, 2010, Pp. 345-357

[22] M. Idu, J.O. Erhabor, and H.M. Efijuemue. "Documentation on Medicinal Plants Sold in Markets in Abeokuta, Nigeria", *Tropical Journal of Pharmaceutical Research,* April 2010; 9 (2), pp. 110-118

[23] J. D. Olowokudejo, A. B. Kadiri, and V.A. Travih. "An Ethnobotanical Survey of Herbal Markets and Medicinal Plants in Lagos State of Nigeria", *Ethnobotanical Leaflets, 2008, 12, pp. 851-865.*

[24] J. K. Rao, J. Suneetha, T.V.V. S. Reddi, and O. A. Kumar. "Ethnomedicine of the Gadabas, a primitive tribe of Visakhapatnam district, Andhra Pradesh", *International Multidisciplinary Research Journal* 2011, 1/2, pp.10-14

[25] World Health Organization, *WHO monographs on selected medicinal plants* volumes 1- 4, WHO Press, World Health Organization, 20 Avenue Appia, 1211 Geneva 27, Switzerland, 2009

[26] Natural Medicines Comprehensive Database, www.naturaldatabase.com

[27] Natural Standard, www.naturalstandard.com

[28] American Botanical Council, www.herbalgram.org

[29] Botanic Garden Conservation International, www.bgci.org

[30] T. Odugbemi, www.medicinalplantsinnigeria.com

[31] ScienceDirect, www.sciencedirect.com

[32] Biomed Central, www.biomedcentral.com

[33] Springerlink, www.springerlink.com

[34]PubMed, www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed

[35] C. Kettner; H. Kosch; M. Lang; J. Lachner; D. Oborny; and E. Teppan. "Creating a Medicinal Plant Database", http://www.beilstein-institut.de/englisch/1024.htm. Accessed January 10, 2011

[36] P.N. Ravindran and K. N. Babu (Eds). *Ginger: the genus Zingiber*, CRC Press, 2005

[37] A.Ali and A.H. Gilani. "Medicinal Value of Ginger With Focus on Its Use in Nausea and Vomiting of Pregnancy", *International Journal of Food Properties*, 2007,10, pp. 269–278

Zacchaeus Oni **Omogbadegun** holds B.Sc (Hons) Computer Science (Second Class Upper Div, 1979) from University of Ibadan, Ibadan, Nigeria and M.Sc Computer Science (2003) from University of Lagos, Lagos, Nigeria. He has over thirty years of progressively cognate experience in Information Technology spanning Software Engineering, Education and Training. He has worked either as an employee or Information Technology Consultant in bluechip industrial organizations within and outside Nigeria: Mobil Oil Nigeria Ltd (1980-1990), Equitorial Trust Bank (1990-1994), and as an Information Technology Consultant to: Ghana's Social Security & National Insurance Trust Software Replacement, Accra, Ghana, (1994-1996); Information Technology Consultant [Contract], FAO of the United Nations, Regional Office for Africa, Accra, Ghana (06/1997 – 09/1997); Edo State Government of Nigeria's Ministry of Education (01/1998- 12/1999); Ondo State Government of Nigeria & UNDP (Sept; 2000); Ekiti State Government of Nigeria's Ministry of Health (Sept; 2001). He has represented his employers in Tanzania, United Kingdom, and Zimbabwe at Software Internationalization Project programmes (1985-1997). He joined academics as Lecturer I (Computer Science) at The Federal

Polytechnic, Ado-Ekiti, Ekiti State, Nigeria (June 2001 – September 2005). He joined Covenant University, Ota, Ogun State, Nigeria as Lecturer II (October 2005) and became Lecturer I (2008 - present). He is currently a PhD (Computer Science) student in the Department of Computer and Information Sciences, Covenant University, Ota, Ogun State, Nigeria. His research interests are Software Engineering (Formal Methods), Healthcare Informatics, MDG's Education and Health-related goals, Artificial Intelligence, Education Informatics, Neural Networks, Biomedical Engineering, Computer Security, Multimedia Database System, and Information Technology Laws. His presentations at both local and international conferences included *"Security in Healthcare Information Systems",* "*Impact of Mobile and Wireless Technologies on Healthcare Delivery Services",* and *"3G and 4G Technologies' Framework for Realising Millenium Development Goals in Healthcare in Nigeria"*, He is a member of the Nigerian Computer Society (NCS), and Computer Professional Registration Council of Nigeria (CPN).

Professor Charles Onuwa **Uwadia** holds a BSc. Degree honours in Computer Science from the University of Ibadan in 1979. He had his MSc. in 1983 and PhD in 1990 both at the University of Lagos. Professor Uwadia joined the services of the University of Lagos as an Assistant Lecturer in 1983, and rose steadily to the post of a full Professor of Computer Science in the year 2004. His major area of specialization is Software Engineering with emphasis on Compiling Techniques and Systems Software. He is also actively involved in teaching and research work in networking, congestion control and management aspects of Information Technology. With over 50 publications in both local and international journals, he has three books to his name.

He is an active member of the Nigeria Computer Society (NCS), and the Computer Professionals Registration Council of Nigeria. He is also a Fellow of NCS. He had served NCS in various capacities including: Chairman Education Committee (1997 – 1999); Chairman Conferences Committee (1999 – 2003); Chairman Publications Committee (2003 – July 2007). He is the current President of the Society. Besides his enviable service in Nigeria, Professor Uwadia has been a visiting Fellow to Universities and Institutions in the United States of America, Europe, Asia, and Australia. He has many publications to his credit. Professor Charles Onuwa Uwadia is the current Director of the Centre for Information Technology and Systems (CITS) of the University of Lagos; a position he has held since 2005.

Professor **Ayo** Charles Korede holds a B.Sc. M.Sc. and Ph.D in Computer Science. He is currently the Director, Academic Planning Unit of Covenant University. He was the pioneer Head of Computer and Information Sciences Department of the University. His research interests include: Mobile computing, Internet programming, eBusiness, eGovernment and Software Engineering. He is a member of the Nigerian Computer Society (NCS), and Computer Professionals (Registration Council) of Nigeria (CPN). Similarly, he is professionally certified in CISCO and Microsoft products. Prof. Ayo is a member of a number of international research bodies such as the Centre for Business Information, Organization and Process Management (BIOPoM), University of Westminster, London; the Review Committee of the European Conference on E-Government ECEG); the programme committee, IADIS Information Systems; the Editorial Board, Journal of Information and communication Technology for Human Development (IJICTHD), the Editorial Board, International Journal of Scientific Research in Education (IJSRE) and the Editorial Board, African Journal of Business Management amongst others. Furthermore, Prof. Ayo is an External Examiner to a number of Nigerian universities at both Undergraduate and Postgraduate levels in Ladoke Akintola University of Technology, Ogbomoso, the Redeemers University, Ogun State, Bells University of Technology, Ota, etc. He has supervised about 200 postgraduate projects at Postgraduate Diploma, Masters and Ph.D levels, and he has several publications in scholarly journals and conferences.

Third Author is a member of the IEEE and the IEEE Computer Society. Do not specify email address here.

Comparison of Conventional and Modern Load Forecasting Techniques Based on Artificial Intelligence and Expert Systems

Engr. Badar Ul Islam
Head, Department of Computer Science & Engineering
NFC-Institute of Engineering & Fertilizer Research, Faisalabad - Pakistan

Abstract

This paper picturesquely depicts the comparison of different methodologies adopted for predicting the load demand and highlights the changing trend and values under new circumstances using latest non analytical soft computing techniques employed in the field of electrical load forecasting. A very clear advocacy about the changing trends from conventional and obsolete to the modern techniques is explained in very simple way. Load forecast has been a central and an integral process in the planning and operation of electric utilities. Many techniques and approaches have been investigated to tackle this problem in the last two decades. These are often different in nature and apply different engineering considerations and economic analysis. Further a clear comparison is also presented between the past standard practices with the current methodology of electrical load demand forecasting. Besides all this, different important points are highlighted which need special attention while doing load forecasting when the environment is competitive and deregulated one.

1.0 INTRODUCTION

Electrical Load Forecasting is the estimation for future load by an industry or utility company. Load forecasting is vitally important for the electric industry in the deregulated economy. A large variety of mathematical methods have been developed for load forecasting. It has many applications including energy purchasing and generation, load switching, contract evaluation, and infrastructure development.

Now a days, development in every sector is a heading at a very rapid pace and in the same pattern, the demand for power is also growing. While speaking about electrical power, it is important to understand that it has three main sectors i.e. generation, transmission and distribution. Electrical power generated by any source is then transmitted through transmission lines at different voltage level and then distributed to different categories of consumers later on. It is not as simple as described in few words but every stage is a complete independent system in itself. Effective load forecasts can help to improve and properly plan these three fields of power systems [1].

Accurate models for electric power load forecasting are essential to the operation and planning of a utility company. Load forecasting helps an electric utility to make important decisions including decisions on purchasing and generating electric power, load switching, and infrastructure development. Load forecasts are extremely important for energy suppliers, ISOs, financial institutions, and other participants in electric energy generation, transmission, distribution, and markets.

Over the past decade, many western nations have begun major structural reforms of their electricity markets. These reforms are aimed at breaking up traditional regional monopolies and replacing them with several generation and distribution utilities that bid to sell or buy electricity through a wholesale market. While the rules of how various wholesale markets operate differ, in each case it is hoped that the end result is a decline in the price of electricity to end users and a price that better reflects the actual costs involved. To successfully operate in these new markets electricity utilities face two complex statistical problems: how to forecast both electricity load and the wholesale spot price of electricity. Failure to implement efficient solutions to these two forecasting problems can directly result in multimillion dollar losses through uninformed trades in the wholesale market.

Load forecasting is however a difficult task. First, because the load series is complex and exhibits several levels of seasonality: the load at a given hour is dependent not only on the load at the previous hour, but also on the load at the same hour on the previous day, and on the load at the same hour on the day with the same denomination in the previous week. Secondly, there are many important exogenous variables that must be considered, especially weather-related variables. It is relatively easy to get forecast with about 10 % mean absolute error; however, the cost of error are so high that research could help to reduce it in a few percent points would be amply justified [2].

2.0 ELECTRICAL LOAD FORECASTING TYPES

The electricity supply industry requires to forecast electricity demand with lead times that range from the short term (a few minutes, hours, or days ahead) to the long term (up to 20 years ahead). Load forecasting has three techniques shown in Figure 2.1:

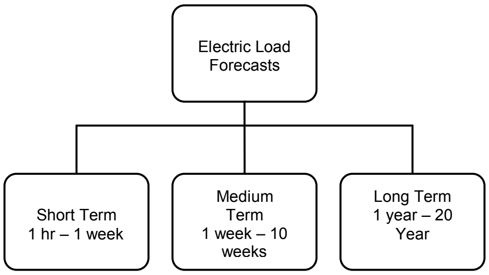


Figure 2.1 Basic Load Forecasting Techniques

Short term electric load forecast spans the period from one hour up to one week and it is mainly utilized for power system operation studies, losses reduction, voltage regulations, unit commitment and maximizing the utility revenues in the deregulated environment. Medium term electric load forecast spans the period from one week to several weeks, it is mainly utilized for predicting the necessary power to purchase or sell from other neighboring networks (inter-tie exchanged power) and also the fuel required by the utility in the near future. In short and medium electric load forecast, it is required to know how much power we will need and at what time of the day; the information regarding where this demand is required is not of a major concern [1].

3.0 ELECTRICAL LOAD FORECASTING METHODS

A model or method is a mathematical description of how the complex elements of a real-life situation or problem might interplay at some future date. In projecting electricity demand, a method uses data on electricity prices, income, population, the economy, and the growth rates for each and then varies the mix according to varying sets of assumptions. Different assumptions produce different outcomes. The relationships between electricity demand and the multitude of factors that influence or affect electricity demand are expressed in mathematical equations called functions. A model is a collection of functions. A function, in turn, is made up of variables for which those factors which change or can be changed. Independent variables are those factors which influence the demand for electricity, and the dependent variable is electricity demand itself. In other words, the demand for electricity depends on population, income, prices, etc. Finally, elasticities describe how much the dependent variable (electricity demand) changes in sense to small changes in the independent variables. Elasticities are what the modeler uses to measure consumer behavior.

Energy planners often speak of scenarios. Hypothetical pictures of the future based on different assumptions about economic or political events. They make different projections for each scenario. For example, a low growth scenario might assume high energy prices and slow population growth, while a high-growth scenario would assume the opposite. These scenarios allow planners to see how electricity demand might change if the different assumed economic and political events actually occur. All of the forecasting methods are capable of looking at different scenarios and do so by changing their basic assumptions [5].

4.0 SHORT TERM LOAD FORECASTING METHODS

Short-Term Load Forecasting is basically is a load predicting system with a leading time of one hour to seven days, which is necessary for adequate scheduling and operation of power systems. It has been an essential component of Energy Management Systems (EMS). For proper and profitable management in electrical utilities, short-term load forecasting has lot of importance.

High forecasting accuracy and speed are the two most important requirements of short-term load forecasting and it is important to analyze the load characteristics and identify the main factors affecting the load. In electricity markets, the traditional load affecting factors such as season, day type and weather, electricity price that have voluntary and may have a complicated relationship with system load..

Various forecasting techniques have been applied to short-term load forecasting to improve accuracy and efficiency. In general, these techniques can be classified as either traditional or modern. Traditional statistical load forecasting techniques, such as regression, time series, pattern recognition, Kalman filters, etc., have been used in practice for a long time, showing the forecasting accuracy that is system dependent. These traditional methods can be combined using weighted multi-model forecasting techniques, showing adequate results in practical systems. However, these methods cannot properly represent the complex nonlinear relationships that exist between the load and a series of factors that influence it, which are typically dependent on system changes (e.g., season or time of day).

The short term load forecasting methods are

- Similar Day Lookup Approach
- Regression Based Approach
- Time Series Analysis
- Artificial Neural Network
- Expert System
- Fuzzy logic
- Support Vector Machines

4.1 Similar Day Look Up Approach

Similar day approach is based on searching historical data of days of one, two or three years having the similar characteristics to the day of forecast. The characteristics include similar weather conditions, similar day of the week or date. The load of the similar day is considered as the forecast. Now, instead of taking a single similar day, forecasting is done through linear combinations or regression procedures by taking several similar days. The trend coefficients of the previous years are extracted from the similar days and forecast of the concern day is done on their basis.

4.2 Regression Based Approach

The term "regression" was used in the nineteenth century to describe a biological phenomenon, namely that the progeny of exceptional individuals tend on average to be less exceptional than their parents and more like their more distant ancestors.

Linear regression is a technique which examines the dependent variable to specified independent. The independent variables are firstly considered because changes occur in them unfortunately. In energy forecasting, the dependent variable is usually demand or price of the electricity because it depends on production which on the other hand depends on the independent variables. Independent variables are usually weather related, such as temperature, humidity or wind speed. Slope coefficients measure the sensitivity of the dependent variable that how they changes with the independent variable. Also, by measuring how significant each independent variable has historically been in its relation to the dependent variable. The future value of the dependent variable can be estimated. Essentially, regression analysis attempts to measure the degree of correlation between the dependent and independent variables, thereby establishing the latter’s predicted values[3].

Regression is the one of most widely used statistical techniques. For electric load forecasting, regression methods are usually used to model the relationship of load consumption and other factors such as weather, day type, and customer class. There are several regression models for the next day peak forecasting. Their models contain deterministic influences such as holidays, random variables influences such as average loads, and exogenous influences such as weather.

4.3 Time Series Analysis

Time series forecasting is based on the idea that

reliable predictions can be achieved by modeling patterns in a time series plot, and then extrapolating those patterns to the future. Using historical data as input, time series analysis fits a model according to seasonality and trend.

Time series models can be accurate in some situations, but are especially complex and require large amounts of historical data. Additionally, careful efforts must made to ensure an accurate time line through out data collection filtering modeling and recall processes. Time series analysis widely used in the martial management for forecasting of customer demand for goods services. Time series approaches are not widely used for energy industry forecasting. Because they typically do not take into account other key factor, such as weather forecasts [3].

Time series have been used for longtime in such fields as economics, digital signal processing, as well as electric load forecasting. In particular, ARMA (autoregressive moving average), ARIMA (autoregressive integrated moving average), ARMAX (autoregressive moving average with exogenous variables), and ARIMAX (autoregressive integrated moving average with exogenous variables) are the most used classical time series methods.

ARMA models are usually used for stationary processes while ARIMA is an extension of ARMA for non-stationary processes. ARMA and ARIMA use the time and load as the only input parameters. Since load generally depends on the weather and time of the day, ARIMAX is the most natural tool for load forecasting among the classical time series models.

4.4 Artificial Neural Networks

Artificial Neural Networks are still at very early stage electronic models based on the neural structure of the brain. We know that the brain basically learns from the experience. The biological inspired methods are thought to be the major advancement in the computational industry. In a neural network, the basic processing element is the neuron. These neurons get input from some source, combine them, perform all necessary operations and put the final results on the output. Artificial neural networks are developed since mid-1980 and extensively applied. They have very successful applications in pattern recognition and many other problems.

Forecasting is based on the pattern observed from the past event and estimates the values for the future. ANN is well suited to forecasting for two reasons. First, it has been demonstrated that ANN are able to approximate numerically any continuous function to be desired accuracy. In this case the ANN is seen as multivariate, nonlinear and nonparametric methods. Secondly, ANNs are date-driven methods, in the sense that it is not necessary for the researcher to use tentative modals and then estimate their parameters. ANNs are able to automatically map the relationship between input and output, they learn this relationship and store this learning into their parameters [3].

The first way is by repeatedly forecasting one hourly load at a time. The second way is by using a system with 24 NNs in parallel, one for each hour of the day. Estimating a model that fits the data so well that it ends by including some of In Multi Layer Perceptron(MLP) structure of neural network, the most commonly training algorithm use is the back propagation algorithm. These algorithms are iterative; some criteria must be defined to stop the iterations. For this either training is stopped after a fixed number of iterations or after the error decreased below some specified tolerance. This criterion is not adequate, this insure that the model fits closely to the training data but does not guarantee of good performance they may lead to over-fitting of the model. "Over-fitting" means the error randomness in its structure, and then produces poor forecasts. MLPs model is over-trained or because it is too complex. One way to avoid overtraining is by using cross-validation. The sample set is split into a training set and a validation set. The neural network parameters are estimated on the training set, and the performance of the model is tested, every few iterations, on the validation set. When this performance starts to deteriorate (which means the neural network is over-fitting the training data), the iterations are stopped, and the last set of parameters to be computed is used to produce the forecasts. Nowadays, other than MLPs to avoid the problems of over-fitting and over-parameterization, the ANNs architectures used for prediction of electrical load are Functional Link Network (FLN) model [1].

To use the ANN in electric load forecast problems, distribution engineers should decide upon a number of basic variables, these variables include:

- Input variable to the ANN (load, temperature…etc)
- Number of classes (weekday, weekend, season…etc)
- What to forecast: hourly loads, next day peak load, next day total load …etc
- Neural network structure (Feedforward, number of hidden layer, number of neuron in the hidden layer…etc)
- Training method and stopping criterion
- Activation functions
- Size of the training data
- Size of the test data

4.5 Expert Systems

Expert systems are new techniques that have come out as a result of advances in the field of artificial intelligence (AI) in the last two decades. An expert system is a computer program, which has the ability to act as an expert. This means this computer program can reason, explain, and have its knowledge base expanded as new information becomes available to it.

The load forecast model is built using the knowledge about the load forecast domain from an expert in the field. The "Knowledge Engineer" extracts this knowledge from load forecast (domain) expert which is called the acquisition module component of the expert system. This knowledge is represented as facts and rules by using the first predicate logic to represent the facts and IF-THEN production rules. This representation is built in what is called the knowledge base component of the expert system. The search for solution or reasoning about the conclusion drawn by the expert system is performed by the "Inference Engine" component of the expert system. For any expert system it has to have the capability to trace its reasoning if asked by the user. This facility is built through an explanatory interface component.

An example demonstrating this approach is the rule-based algorithm which is based on the work of two scientists Rahman and Baba. This algorithm consists of functions that have been developed for the load forecast model based on the logical and syntactical relationship between the weather and prevailing daily load shapes in the form of rules in a rule-base. The rule-base developed consists of the set of relationships between the changes in the system load and changes in natural and forced condition factors that affect the use of electricity. The extraction of these rules was done off-line, and was dependent on the operator experience and observations by the authors in most cases. Statistical packages were used to support or reject some of the possible relationships that have been observed

The rule-base consisted of all rules taking the IF-THEN form and mathematical expressions. This rule-base is used daily to generate the forecasts. Some of the rules do not change over time, some change very slowly while others change continuously and hence are to be updated from time to time [4].

4.6 Fuzzy Logic

Fuzzy logic based on the usual Boolean logic which is used for digital circuit design. In Boolean logic, the input may be the truth value in the form of “0” and “1”. In case of fuzzy logic, the input is related to the comparison based on qualities. For example, we can say that a transformer load may be “low” and “high”. Fuzzy logic allows us to deduce outputs form inputs logically. In this sense, the fuzzy facilitate for mapping between inputs and outputs like curve fitting *[16]*.

The advantage of fuzzy logic is that there is no need of mathematical models for mapping between inputs and outputs and also there is no need of precise or even noise free inputs. Based on the general rules, properly designed fuzzy logic systems are very strong for the electrical load forecasting. There are many situations where we require the precise outputs. After the whole processing is done using the fuzzy logic, the "defuzzification" is done to get the precise outputs.

We know that power system load is influenced by many load factors such weather, economic and social activities and different load components. By the analysis of historical load data it is not easy to make the accurate forecast. The use of these intelligent methods like fuzzy logic and expert systems provide advantage on other conventional methods. The numerical aspects and uncertainties are suitable for the fuzzy methodologies[5].

4.7 Support Vector Machines

Support Vector Machines (SVM) are the most powerful and very recent techniques for the solution of classification and regression problems. This approach was come to known from the work of Vapnik's, his statistical learning theory. Other from the neural network and other intelligent systems, which try to define the complex functions of the inputs, support vector machines use the nonlinear mapping of the data in to high dimensional features by using the kernel functions mostly. In support vector machines, we use simple linear functions to create linear decision boundaries in the new space. In the case of neural network, the problem is in the choosing of architecture and in the case of support vector machine, problems occurs in choosing a suitable kernel.

Mohandes applied a method of support vector machines for short-term electrical load forecasting. He compares its method performance with the autoregressive method. The results indicate that SVMs compare favorably against the autoregressive method. Chen also proposed a SVM model to predict daily load demand of a month. Lots of methods are used in support vector machines [3].

5.0 MEDIUM AND LONG-TERM LOAD FORECASTING METHODS

These models are useful for medium and long term forecasting. The three types of electricity demand forecasting methods are:

1. Trend Analysis
2. End Use Analysis
3. Econometrics

Each of the three forecasting methods uses a different approach to determine electricity demand during a specific year in a particular place. Each forecasting method is distinctive in its handling of the four basic forecast ingredients:

1. The mathematical expressions of the relationship between electricity demand and the factors which influence or affect it - the function
2. The factors which actually influence electricity demand (population, income, prices, etc.) - the independent variables
3. Electricity demand itself - the dependent variable
4. How much electricity demand changes in response to population, income, price, etc., changes- the elasticities?

The only way to determine the accuracy of any load forecast is to wait until the forecast year has ended and then compare the actual load to the forecast load. Even though the whole idea of forecasts is accuracy, nothing was said in the comparison of the three forecasting methods about which method produces the most accurate forecasts. The only thing certain shut any long-range forecast is that it can never be absolutely precise. Forecasting accuracy depends on the quality and quantity of the historical data used, the validity of the forecasters basic assumptions, and the accuracy of the forecasts of the demand-influencing factors (population, income, price, etc.). None of these is ever perfect. Consequently, regional load forecasts are reviewed some are revised yearly. Even so, there is simply electricity demand will be exactly as forecast, no is used or who makes the forecast. Continually, and no assurance that matter what method is used or who makes the forecast [3].

5.1 Trend Analysis

Trend analysis (trending) extends past growth rates of electricity demand into the future, using techniques that range from hand-drawn straight lines to complex computer-produced curves. These extensions constitute the forecast. Trend analysis focuses on past changes or movements in electricity demand and uses them to predict future changes in electricity demand. Usually, there is not much explanation of why demand acts as it does, in the past or in the future. Trending is frequently modified by informed judgment, wherein utility forecasters modify their forecasts based on their knowledge of future developments which might make future electricity demand behave differently than it has in the past.

The advantage of trend analysis is that it is simple, quick and inexpensive to perform. It is useful when there is not enough data to use more sophisticated methods or when time and funding do not allow for a more elaborate approach.

The disadvantage of a trend forecast is that it produces only one result - future electricity demand. It does not help analyze why electricity demand behaves the way it does, and it provides no means to accurately measure how changes in energy prices or government policies (for instance) influence electricity demand. Because the assumptions used to make the forecast (informed judgments) are usually not spelled out, there is often no way to measure the impact of a change in one of the assumptions. Another shortcoming of trend analysis is that it relies on past patterns of electricity demand to project future patterns of electricity demand. This simplified view of electrical energy could lead to inaccurate forecasts in times of change, especially when new concepts such as conservation and load management must be included in the analysis [3].

5.2 End Use Analysis

The basic idea of end-use analysis is that the demand for electricity depends on what it is used for (the end-use). For instance, by studying historical data to find out how much electricity is used for individual electrical appliances in homes, then multiplying that number by the projected number of appliances in each home and multiplying again by the projected number of homes, an estimate of how much electricity will be needed to run all household appliances in a geographical area during any particular year in the future can be determined. Using similar techniques for electricity used in business and industry, and then adding up the totals for residential, commercial, and industrial sectors, a total forecast of electricity demand can be derived. The advantages of end-use analysis is that it identifies exactly where electricity goes, how much is used for each purpose, and the potential for additional conservation for each end-use. End-use analysis provides specific information on how energy requirements can be reduced over time from conservation measures such as improved insulation levels, increased use of storm windows, building code changes, or improved appliance efficiencies. An end-use model also breaks down electricity into residential, commercial and industrial demands. Such a model can be used to forecast load changes caused by changes within one sector (residential, for example) and load changes resulting indirectly from changes in the other two sectors. Commercial sector end-use models currently being developed have the capability of making energy demand forecasts by end-uses as specific as type of business and type of building. This is a major improvement over projecting only sector-wide energy consumption and using economic and demographic data for large geographical areas [1].

The disadvantage of end-use analysis is that most end-use models assume a constant relationship between electricity and end-use (electricity per appliance, or electricity used per dollar of industrial output). This might hold true over a few years, but over a 10-or 20-year period, energy savings technology or energy prices will undoubtedly change, and the relationships will not remain constant. End-use analysis also requires extensive data, since all relationships between electric load and all the many end-uses must be calculated as precisely as possible. Data on the existing stock of energy-consuming capital (buildings, machinery, etc.) in many cases is very limited. Also, if the data needed for end-use analysis is not current, it may not accurately reflect either present or future conditions, and this can affect the accuracy of the forecast. Finally, end-use analysis, without an econometric component that is explained above, does not take price changes (elasticity of demand) in electricity or other competing fuels into consideration.

Ideally this approach is very accurate. However, it is sensitive to the amount and quality of end-use data. For example, in this method the distribution of equipment age is important for particular types of appliances. End-use forecast requires less historical data but more information about customers and their equipment [1].

5.3 Econometric

Econometrics uses economics, mathematics, and statistics to forecast electricity demand. Econometrics is a combination of trend analysis and end-use analysis, but it does not make the trend-analyst’s assumption that future electricity demand can be projected based on past demand. Moreover, unlike many end-use models, econometrics can allow for variations in the relationship between electricity input and end-use.

Econometrics uses complex mathematical equations to show past relationships between electricity demand and the factors which influence that demand. For instance, an equation can show how electricity demand in the past reacted to population growth, price changes, etc. For each influencing factor, the equation can show whether the factor caused an increase or decrease in electricity demand, as well as the size (in percent) of the increase or decrease. For price changes, the equation can also show how long it took consumers to respond to the changes. The equation is then tested and fine tuned to make sure that it is as reliable a representation as possible of the past relationships. Once this is done, projected values of demand-influencing factors (population, income, prices) are put into the equation to make the forecast. A similar procedure is followed for all of the equations in the model.

The advantages of econometrics are that it provides detailed information on future levels of electricity demand, why future electricity demand increases or decreases, and how electricity demand is affected by various factors. In addition, it provides separate load forecasts for residential, commercial, and industrial sectors. Because the econometric model is defined in terms of a multitude of factors (policy factors, price factors, end-use factors), it is flexible and useful for analyzing load growth under different scenarios.

A disadvantage of econometric forecasting is that in order for an econometric forecast to be accurate, the changes in electricity demand caused by changes in the factors influencing that demand must remain the same in the forecast period as in the past. This assumption (which is called constant elasticities) may be hard to justify, especially where very large electricity price changes (as opposed to small, gradual changes) make consumers more sensitive to electricity prices *[3]*.

Also, the econometric load forecast can only be as accurate as the forecasts of factors which influence demand. Because the future is not known, projections of very important demand-influencing factors such as electricity, natural gas, or oil prices over a 10- or 20-year period are, at best, educated guesses. Finally) many of the demand-influencing factors which may be treated and projected individually in the mathematical equations could actually depend on each other, and it is difficult to determine the nature of these interrelationships. For example, higher industrial electricity rates may decrease industrial employment, and projecting both of them to increase at the same time may be incorrect. A model which treats projected industrial electricity rates and industrial employment separately would not show this fact.

Econometric models work best when forecasting at national, regional, or state levels. For smaller geographical areas, meeting the model can be a problem. This is oddly shaped service areas for which there demographic data.

6.0 COMPARISON OF ELECTRICAL LOAD FORECASTING TECHNIQUES

In the previous discussion we focus on electrical load forecasting techniques, most forecasting methods use statistical techniques or artificial intelligence algorithms such as regression, neural networks, fuzzy logic, and expert systems. Two of the methods named trend analysis, end-use and econometric approach are broadly used for medium- and long-term forecasting. A variety of methods, which include the similar day approach, various regression models, time series, neural networks, statistical learning algorithms, fuzzy logic, and expert systems, have been developed for short-term forecasting.

The method for short-term forecasting are similar day approach, various regression models, time series, neural networks, statistical learning algorithms, fuzzy logic, and expert systems. ***Similar day approach*** is based on searching historical data of days of one, two or three years having the similar characteristics to the day of forecast. ***Regression*** is the one of most widely used statistical techniques. For electric load forecasting, regression methods are usually used to model the relationship of load consumption and other factors such as weather, day type, and customer class. There are several regression models for the next day peak forecasting. Their models contain deterministic influences such as holidays, random variables influences such as average loads, and exogenous influences such as weather. ***Time series*** is a very popular approach for the electrical load forecasting. Two important models of time series are ARMA and ARIMA. ARMA and ARIMA use the time and load as the only input parameters. Since load generally depends on the weather and time of the day, ARIMAX is the most natural tool for load forecasting among the classical time series models *[2]*.

The other methods are based on Artificial intelligence, they are called Intelligent Systems. In ***Artificial Neural Network***, forecasting is based on the pattern observed from the past event and estimates the values for the future. ANN is well suited to forecasting for two reasons. First, it has been demonstrated that ANN are able to approximate numerically any continuous function to be desired accuracy. In this case the ANN is seen as multivariate, nonlinear and nonparametric methods. Secondly, ANNs are date-driven methods, in the sense that it is not necessary for the researcher to use tentative modals and then estimate their parameters. ANNs are able to automatically map the relationship between input and output, they learn this relationship and store this learning into their parameters. An ***Expert System*** is a computer program, which has the ability to act as an expert. This means this computer program can reason, explain, and have its knowledge base expanded as new information becomes available to it. The load forecast model is built using the knowledge about the load forecast domain from an expert in the field. This knowledge is represented as facts and rules by using the first predicate logic to represent the facts and IF-THEN production rules. This representation is built in what is called the knowledge base component of the expert system. The search for solution or reasoning about the conclusion drawn by the expert system is performed by the "Inference Engine" component of the expert system. For any expert system it has to have the capability to trace its reasoning if asked by the user. This facility is built through an explanatory interface component. ***Fuzzy logic*** based on the usual Boolean logic which is used for digital circuit design. In case of fuzzy logic, the input is related to the comparison based on qualities. The advantage of fuzzy logic is that there is no need of mathematical models for mapping between inputs and outputs and also there is no need of precise or even noise free inputs. Based on the general rules, properly designed fuzzy logic systems are very strong for the electrical load forecasting.

The methods for long- and medium-term forecasting are trend analysis, end-use and econometric approach. The advantage of ***trend analysis*** is that it is quick, simple and inexpensive to perform and does not require much previous data. The basic idea of the ***end-use analysis*** is that the demand for electricity depends what it use for (the end-use). The advantages of end-use analysis is that it identifies exactly where electricity goes, how much is used for each purpose, and the potential for additional conservation for each end-use. The disadvantage of end-use analysis is that most end-use models assume a constant relationship between electricity and end-use (electricity per appliance, or electricity used per dollar of industrial output). This might hold true over a few years, but over a 10-or 20-year period, energy savings technology or energy prices will undoubtedly change, and the relationships will not remain constant. The advantages of ***econometrics*** are that it provides detailed information on future levels of electricity demand, why future electricity demand increases or decreases, and how electricity demand is affected by various factors. A disadvantage of econometric forecasting is that in order for an econometric forecast to be accurate, the changes in electricity demand caused by changes in the factors influencing that demand must remain the same in the forecast period as in the past *[5]*.

7.0 CONCLUSION

Modern load forecasting techniques, such as expert systems, Artificial Neural Networks (ANN), fuzzy logic, wavelets, have been developed recently, showing encouraging results. Among them, ANN methods are particularly attractive, as they have the ability to handle the nonlinear relationships between load and the factors affecting it directly from historical data.

The trend analysis, end-use modeling and econometric modeling are the most often used methods for medium- and long-term load forecasting. Trend analysis (trending) extends past growth rates of electricity demand into the future, using techniques that range from hand-drawn straight lines to complex computer-produced curves. Descriptions of appliances used by customers, the sizes of the houses, the age of equipment, technology changes, customer behavior, and population dynamics are usually included in the statistical and simulation models based on the so-called end-use approach. In addition, economic factors such as per capita incomes, employment levels, and electricity prices are included in econometric models. These models are often used in combination with the end-use approach. Long-term forecasts include the forecasts on the population changes, economic development, industrial construction, and technology development.

REFERENCES

[1] "Computational Intelligence in Time Series Forecasting Theory and Engineering Applications" (Advances in Industrial Control) by: Ajoy K. Palit, Dobrivoje Popovic, Springer, 2005.

[2] Ibrahim Mogharm , Saifur Rehaman, "Analysis and Evaluation of Five Short Term Load Forecasting Techniques" , IEEE Transactions on Power Systems, Vol. 4 No. 4, October 1989.

[3] H.L.Willis, "Distribution load forecasting", IEEE Tutorial course on power distribution planning, EHO 361-6-PWR, 1992.

[4] J.V. Ringwood "Intelligent Forecasting of Electricity Demand". www.forecastingprinciples.com

[5] Andrew P. Douglas, Arthur M. Breipohl, "Risk Due To Load Forecast Uncertainty in Short Term Power System Planning" IEEE Transactions on Power Systems, Vol. 13 No. 4, November, 1998.

A New Approach to Encoding and Hiding Information in an Image

Dr. Fadhil Salman Abed
Technical Institute of Kalar
Diyala,Iraq

Abstract

The information age brings some unique challenges to society. New technology and new applications bring new threats and force us to invent new protection mechanisms. So every few years, computer security needs to reinvent itself. In this paper we propose a new image encoding system utilizing fractal theories; this approach exploits the main feature of fractals generated by IFS techniques. Two levels of encryption and decryption methods performed to enhance the security of the system, this is based on the fact that all fractal functions use real number to ensure satisfaction of contraction property. If the cryptosystem parameters are based on real numbers (a continuous infinite interval) then the search space is massive. Hence, many well known attacks fail to solve the nonlinear systems and find the imprecise secret key parameter from the given public one. Even if it is theoretically possible, it is computationally not feasible. The encrypted date represents the attractor generated by the IFS transformation, Collage theorem is used to find the IFS for decrypting data. The proposed method gives the possibility to hide maximum amount of data in an image that represent the attractor of the IFS without degrading its quality.

Also to make the hidden data robust enough to withstand known cryptographic attacks and image processing techniques which do not change the appearance of image. The security level is high because the jointly coded images cannot be correctly reconstructed without all the required information.

Keywords: Image processing , Hiding, Fractal Image Compression, Quadtree, Steganography

1.0 Introduction

Cryptography is the study of mathematical and computational techniques related to aspects of information security. It is a method of transferring private information and data through open network communication, so only the receiver who has the secret key can read the encrypted messages which might be documents, phone conversations, images or other form of data. Modern telecommunication networks, and especially the internet and mobile-phone networks, have tremendously extended the limits and possibilities of communications and information transmissions. Associated with this rapid development, there is a growing demand of cryptographic techniques, which has spurred a great deal of intensive research activities in the study of cryptography. To implement privacy simply by encrypting the information intended to remain secret can be achieved by using methods of cryptography. The information must be scrambled, so that other users will not be able to access the actual information. For example, in a multi-users system, each user may keep his privacy intact via her/his own password. On internet, a large number of internet users use internet application, such as business, research, learning, etc. These activities are very important for the users, application; hence, the importance of using cryptography has been highlighted to help them keep the privacy **[1]**.

Since 1990s, many researchers have noticed that there exists an interesting relationship between chaos, fractal and cryptography. Dynamical systems theory is closely related to fractal geometry. One can show that fractals attractors of iterated function systems in particular have a naturally associated dynamical system which is chaotic. Fractals are attractors of dynamical systems; the place where chaotic dynamics occur. Many properties of chaotic systems have their corresponding counter parts in traditional cryptosystems; they are characterized by sensitive dependence on initial conditions, similarity to random behavior, and continuous broad-band power spectrum. The suggested guidelines address three main issues: implementation, key management, and security analysis, aiming at assisting designers of new cryptosystems to present their work in a more systematic and rigorous way to fulfill some basic cryptographic requirements. In recent years, a large amount of work on chaos-based cryptosystems has been published. Much work has been done by incorporating chaotic maps into the design of symmetric and asymmetric encryption scheme. In 2003, Kocarev and Tasev proposed a public key encryption algorithm based on chebyshev chaotic maps, and after that many works that proposed a new key agreement protocol based on chaotic maps are developed. Also some works for incorporating of fractal functions into the design of symmetric and asymmetric encryption schemes using the similar mechanism have been proposed in **[2]**.

The use of fractal have advantage since; only few parameter would have to be stored, and this kind of key is very robust to attacks for these two reasons; if the attacker managed to obtain parts of the key (or almost the entire key), but a small digit is missing or is incorrect, the fractal image is changed dramatically. In this case the attacker has no way to extrapolate the rest of the key. The second reason, the brute force attack will not work since a fractal key is time consuming to generate especially at high zoon levels. Fractal geometry and, in particular, the theory of fractal functions, has evolved beyond its mathematical framework and has become a powerful and useful tool in the applied sciences as well as engineering. The realm of applications includes structural mechanics, physics and chemistry, signal processing and decoding, and cryptography. The reason for this variety of applications lies in the underlying complicated mathematical structure of fractal functions, specifically their recursive construction. For certain problems they provide better approximants than their classical non-recursive counterparts**[3]**.

2.0 The Proposed Encoding and Hiding System

The main objects of the proposed system contain two stage, firstly by cryptography(encryption) the information massage by using new approach in cryptography based on iterated function system(IFS), secondly by hiding the encryption information by using proposed technique based on fractal image compression.

2.1 Proposed Approaches(Cryptography Units)

There are many types of cryptography in which there are "double enciphering" and "double deciphering" processes that make the codes more difficult to crack and to analyses. The proposed approach for enciphering and deciphering apply two level method for each, for enciphering, firstly, one, by arranging the resulting code in a chosen manor of affine IFS transformation, and the resulting enciphering code is the attractor of the IFS system, secondly by hiding the enciphering text in an image by using Fractal Image Compression.

<u>**Theorem**</u>

$\beta(\mathrm{X})=AX+b$ could be used as a secret key to encipher p messages of length m at a time in n-letter alphabet if and only if GCD(D, n^m)=1.

Proof:-

If B is secret key then B is one to one map from Z_t to Z_t where $\boldsymbol{t= n^m}$ and hence onto and so invertible.
Thus GCD(D, n^m)=1. Conversely if GCD(D, n^m)=1, then A is invertible and hence β is one to one.□

The sender arranges each unit of length m in entries with value one in the affine IFS transformation. The elements of the B maps are constructed from $\boldsymbol{(C_{ij}/n^m)}$ where $\boldsymbol{C_{ij}=p_1n^m+p_2n^{m-1}+\ldots+p_m}$.

2.1.1 Affine IFS maps

An IFS is a standard way to model natural objects. The intuitive key for deriving IFS that models any given object is self-tiling (similarity). One can always view an object as the union of several an objects. Let the sub-objects be actually scaled-down copies of the original object. Each of these subjects is called a tile. In particular, each sub-object is obtained by applying an affine transformation to the entire object**[4]**.

Now consider the original object with two or more affine transformed copies of itself. The tiling scheme should completely cover the object, even if this necessitates overlapping the tiles. Each transformation used to “create” a tile corresponds exactly to one map in the IFS. In order to create an IFS, one first specifies a finite set of contractive affine transformations {$\boldsymbol{b_i}$; $\boldsymbol{i=1,\dots,n}$} in R^2. In general, a contractive affine transformation β in R^2 is of the form: β(X) = AX + b, which could be used as a secret key to produce an enciphering code. There are different possibilities to arrange element in IFS invertible maps, therefore, for abbreviation, binary sequences of 0’s and 1’s used to represent all possibilities for element arranging in the $\boldsymbol{b_i}$ maps, as follows:

$$\beta\begin{bmatrix}x\\y\end{bmatrix}=\begin{pmatrix}a_{11}&a_{12}\\a_{21}&0\end{pmatrix}\begin{bmatrix}x\\y\end{bmatrix}+\begin{bmatrix}0\\0\end{bmatrix}=AX\rightarrow 111000$$

$$\beta\begin{bmatrix}x\\y\end{bmatrix}=\begin{pmatrix}a_{11}&0\\a_{21}&a_{22}\end{pmatrix}\begin{bmatrix}x\\y\end{bmatrix}+\begin{bmatrix}0\\0\end{bmatrix}=AX\rightarrow 101100$$

$$\beta\begin{bmatrix}x\\y\end{bmatrix}=\begin{pmatrix}a_{11}&0\\0&a_{22}\end{pmatrix}\begin{bmatrix}x\\y\end{bmatrix}+\begin{bmatrix}0\\0\end{bmatrix}=AX\rightarrow 100100$$

$$\beta\begin{bmatrix}x\\y\end{bmatrix}=\begin{pmatrix}a_{11}&a_{12}\\a_{21}&a_{22}\end{pmatrix}\begin{bmatrix}x\\y\end{bmatrix}+\begin{bmatrix}0\\0\end{bmatrix}=AX\rightarrow 111100$$

$$\beta\begin{bmatrix}x\\y\end{bmatrix}=\begin{pmatrix}0&a_{12}\\a_{21}&0\end{pmatrix}\begin{bmatrix}x\\y\end{bmatrix}+\begin{bmatrix}0\\0\end{bmatrix}=AX\rightarrow 110000$$

All the above orders are for linear affine transformation. Now for non-linearity order each one of the above maps is extended to three forms by adding the translation part ***b***. For example, for β=111000, we have:

$$\beta\begin{bmatrix}x\\y\end{bmatrix}=A\begin{bmatrix}x\\y\end{bmatrix}+\begin{bmatrix}e\\0\end{bmatrix}=AX+b\rightarrow 111010$$

$$\beta\begin{bmatrix}x\\y\end{bmatrix}=A\begin{bmatrix}x\\y\end{bmatrix}+\begin{bmatrix}e\\f\end{bmatrix}=AX+b\rightarrow 111011$$

$$\beta\begin{bmatrix}x\\y\end{bmatrix}=A\begin{bmatrix}x\\y\end{bmatrix}+\begin{bmatrix}0\\f\end{bmatrix}=AX+b\rightarrow 111001$$

2.1.2 The Implementation

Conversion of the plain-text message to the unreadable format is known as enciphering of the message. Similarly, conversion of the enciphered message back to the human readable form through the reversal of the encryption algorithm is known as deciphering of the message.

2.1.3 Encryption method

Let's assume that there are two parties(sender and receiver) in two far places that need to communicate secretly in a way that a third person (intruder) won't figure or recognize that they are exchanging information between them. However, the alphabetic, the classical encryption method and the order of the affine IFS maps must be agreed upon between sender and receiver.

2.1.4 Enciphering algorithm

In this algorithm an alphabet of n = 29 character is chosen:

- The message characters are given a numbers as it appear in **Table(1)**, show the length of the message.
- Divide the message of length l into units of length m = 3, represented by $\mathbf{p_i p_{i+1} p_{i+2}}$
- Calculate the numeric value of each unit using the polynomial $\mathbf{C = p_i\, n^2+p_{i+1}n+p_{i+2}}$, or matrices operation to perform first level of the proposed method.
- The contraction factor used is $\mathbf{r = 1/n^m}$
- The elements of the chosen affine IFS transformations βi are calculated by $\boldsymbol{\beta_i = r*C}$.

  Notice that B = {β_1, β_2,… β_i} called a (hyperbolic) IFS.
- The attractor A is generated using Random Iterated Algorithm**[1].**

```
(1)  Initialize x=0, y=0 (Starting point).
(2)  Choose arbitrary k  to be one of the
     numbers 1 , 2, 3 , . . . .n , with probability $p_k$ .
(3)  Apply the transformation $w_k$ on the
     point (x, y) to obtain the point (x', y').
(4)  Plot the point (x', y').
(5)  Set x = x' and y = y'.
(6)  Goto step 2.
```

- The enciphering code is the picture represents the Attractor A.

Algorithm (1) Image generation with random IFS

Table (1): English alphabet used for encryption

2.1.5 Decryption Algorithm

- Upon the receipt of the attractor (picture) *A*, the receiver retrieves *B* using "Inverse Problems" techniques. Let A denote the image we want to encode. Let also A_r denote a partition of *A* in nxn blocks referred to as Range blocks (R_b). Similarly, A_d will denote another partition of *A*, this time in 2nx2n blocks or Domain blocks (D_b) in steps of nxn pixels.
- The goal of the encoding algorithm is to establish a relationship between A_r and A_d in such a way that any R_b can be expressed as a set of transformations to be applied on a particular D_b. The receiver then modifies the entries of the retrieved IFS system *B* to get β_i as they agreed on.
- By multiplying each entry in the affine IFS map by n^m and rounding them to the nearest integer we perform the first level of decrypting method.
- Finally Apply some algebraic calculation to find p_1,p_2,p_3 in each cipher unit, as follows.

  p_1=int(C/n^2), R= C mod n^2 ,p_2=int(R/n) ,p_3=R mod n

Example: To encrypt the message, "**We must be good in cipher system.**", the sender and the receiver agreed on an alphabet mentioned in Table 1. The message is divided into units of three characters and used as inputs to the affine transformations after applying the polynomial C = p_i n^2+p_{i+1}n+p_{i+2}, the enciphering code is shown in **Table (2)**. If the affine mappings, 111001, 101110, 111000, 100111 are chosen, then the IFS are constructed as follows:

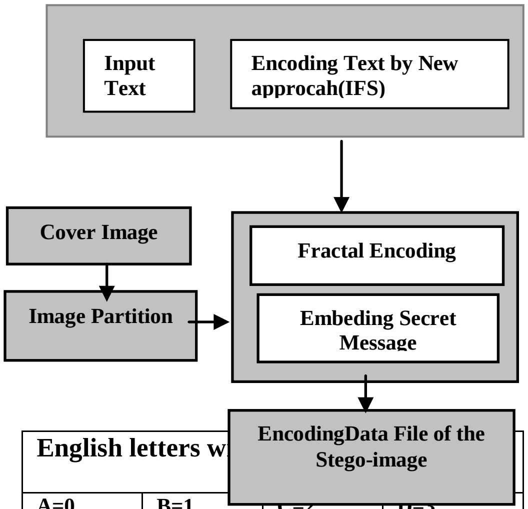
$$B=\bigcup\begin{cases}\frac{1}{29^3}\begin{pmatrix}18644 & 10690\\16734 & 0\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}+\begin{pmatrix}0\\4124\end{pmatrix}, & \text{Prop.}=.1\\[2ex]\frac{1}{29^3}\begin{pmatrix}12183 & 0\\2211 & 21932\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}+\begin{pmatrix}12822\\0\end{pmatrix}, & \text{Prop.}=.2\\[2ex]\frac{1}{29^3}\begin{pmatrix}15068 & 20725\\3738 & 0\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}, & \text{Prop.}=.7\end{cases}$$

Message units	Value	Message units	Value
We$	18644	$ci	21932
Mus	10960	phe	12822
T$b	16734	R$s	15068
E$g	4124	yst	20725
ood	12183	Em.	3739
$in	22111	-	-

Table (2): Message units and their enciphering code

2.2 The Proposed Hiding System(hiding Units)

Figure (1) describes the proposed system starting with loading cover image, then performing the quad-tree partition and then hiding module by starting search process for similarity between the image blocks (range blocks and domain blocks) and embedding the secret message in the scaling and offset values of the blocks. The output of this stage is a data file of stego-image which is sent from a sender to a recipient. When the recipient receives the data file of the stego-image, the process of extraction could be applied to obtain the secret message with an approximate or the reconstructed image.



English letters w				
A=0	B=1	C=2	D=3	
E=4	F=5	G=6	H=7	
I=8	J=9	K=10	L=11	
M=12	N=13	O=14	P=15	
Q=16	R=17	S=18	T=19	
U=20	V=21	W=22	X=23	
Y=24	Z=25	$=26	.=27	?=28

Figure (1): Embedding information unit

The extraction stage illustrated in **figure (2)** starts by loading the file of the stegoimage and extracting the hidden information that received with fractal decoding side by side. Fractal decoding starts by setting all the domains to arbitrary shapes, it goes then into loop. The first iteration applies the transformation to domains that all are black. This creates range that may already after this single iteration, slightly resembles the original ranges. With recreating every block of the ranges the hidden characters will be extracted and the receiver will receive the secret message.

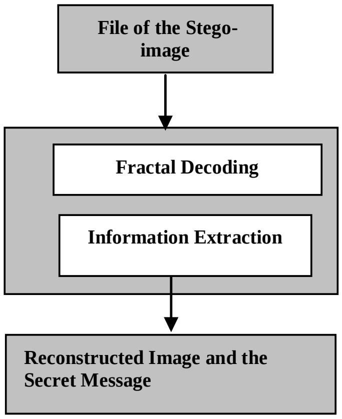


Figure (2): Information extraction unit

2.2.1 Hiding Unit

The structure of the hiding unit; mainly it consists of eight modules:

- Loading cover image.
- Colour separation.
- Convert the image formula from RGB to YCbCr **[5]**

$Y = (77/256)\ R + (150/256)\ G + (29/256)\ B$
$Cb = -(44/256)\ R - (87/256)\ G + (131/256)\ B + 128$
$Cr = (131/256)\ R - (110/256)\ G - (21/256)\ B + 128$

- Partitioning the cover image by using Quad-tree colour image Partition
- Down sampling.
- Fractal encoding.
- Embedding the secret message.
- Stegoimage data file saving.

2.2.2 Quad-tree Partition

One of the most familiar partition techniques is the quadtree method, which subdivides a region of an image into four equal blocks when a given homogeneity criterion is not met by that region. It continues to divide each sub-division until the criteria is met or minimum block size is reached. Typically, an image is initially divided into a set of large blocks (their size equal to the maximum allowable block size). The variance is computed and compared to a threshold for each of these blocks. Any sub-blocks created by failure of the homogeneity test undergo the same procedure. The subdivision will continue until a block either reaches a minimum size or it satisfies the homogeneity criterion. Each block test constitutes a node of the quadtree. A node for which no further subdivision is needed is called a leaf. The tree structure and accompanying encoding for each leaf node are stored or transmitted for later reconstruction. **Figure (3)** illustrates the quadtree partitioning procedure.

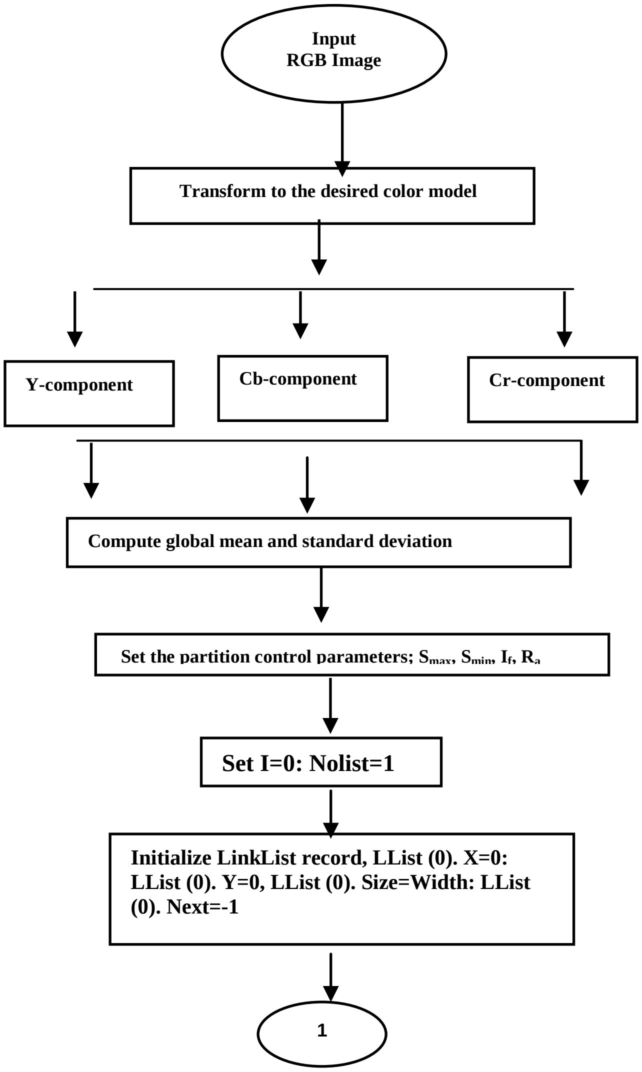
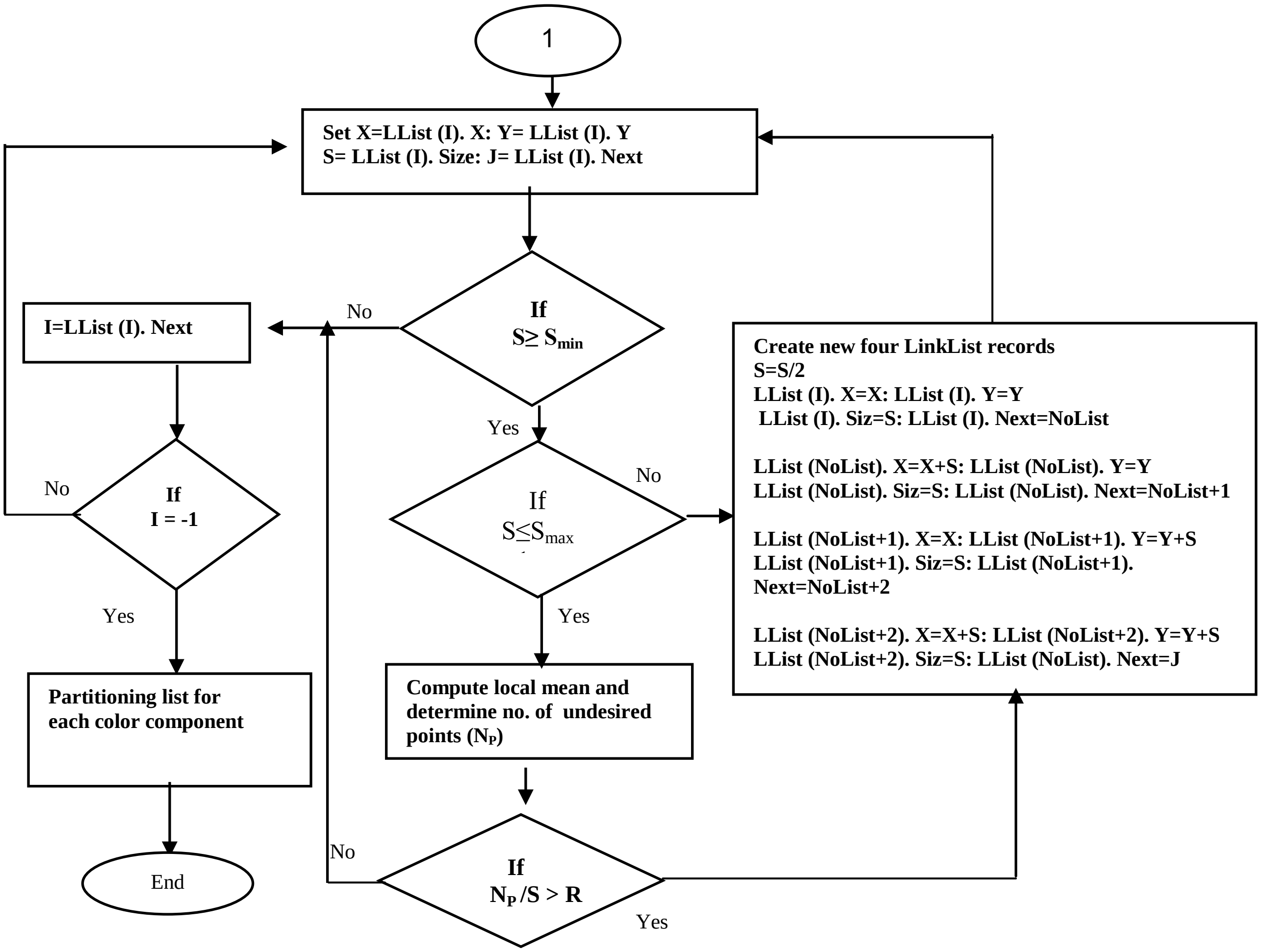

Input RGB Image
Transform to the desired color model
Y-component
Cb-component
Cr-component
Compute global mean and standard deviation
Set the partition control parameters; S_max, S_min, I_f, R_a
Set I=0: Nolist=1
Initialize LinkList record, LList (0). X=0: LList (0). Y=0, LList (0). Size=Width: LList (0). Next=-1
1

Figure (3): Illustrates the quadtree partitioning procedure

For more details about fractal image compression[**[6], [7], [8].**

2.2.4 Secret Message Embedding

For embedding the secret massage, a new method has proposed and in this a new method a huge number of characters can be embedded, beside the characters or the secret message there may not necessary be a text, it may consist of equation or numbers with another language. After the IFS mapping is coming to the end of the last block and the parameter values


1
Set X=LList (I). X: Y= LList (I). Y
S= LList (I). Size: J= LList (I). Next
If
S≥ S_min
No
I=LList (I). Next
Yes
If
I = -1
No
Yes
Partitioning list for each color component
End
If
S≤S_max
No
Yes
Create new four LinkList records
S=S/2
LList (I). X=X: LList (I). Y=Y
LList (I). Siz=S: LList (I). Next=NoList
LList (NoList). X=X+S: LList (NoList). Y=Y
LList (NoList). Siz=S: LList (NoList). Next=NoList+1
LList (NoList+1). X=X: LList (NoList+1). Y=Y+S
LList (NoList+1). Siz=S: LList (NoList+1). Next=NoList+2
LList (NoList+2). X=X+S: LList (NoList+2). Y=Y+S
LList (NoList+2). Siz=S: LList (NoList). Next=J
Compute local mean and determine no. of undesired points (N_P)
If
N_P /S > R
No
Yes

Continuation of figure (3)

2.2.3 Fractal Image Compression

have been set, the process of embedding starts by reading the inputted secret message and converting it to its binary representation then store them in a new array separately**[9].**

Next, the secret message (SecData) characters are taken one by one and they are converted to its ASCII representation. The length of the secret message is limited to the number of blocks. As the number of blocks increase in tern more characters can be embed, in other words this embedding method depends upon the number of blocks.

$$\text{Length of secret message (SecData)} = (\text{No. of blocks} \times 2) - 1$$

Each character is embedded in the scaling value of the blocks by taking the integer part of the scaling using this equation

$$V = IFS(I).\ Scl - Fix\ (IFS(I).Scl)$$

Then taking the two digit of the fraction part (2 digit after the decimal point) of the value and neglecting the other digits

$$IntFraction = fix(\ V \times 100)$$

$$Fraction = IntFraction \times 0.01$$

Now the secret message will be added to the fraction value in order to occupy the **3rd, 4th** and **5th** places after the decimal point without affecting the **1st** and **2nd** places of the original fraction.

$$IFS(I).\ Scl = Fix(IFS(I).Scl) + Fraction + SecData(I) \times 0.00001$$

The output is a set of scale and offset values that contain the secret massage (SecData) is the last stage in hiding unit

2.2.5 Information Extraction Unit

Extraction units are arranged in a reverse order to the hiding unit. It consists of these stages as illustrates in **figure (4)**

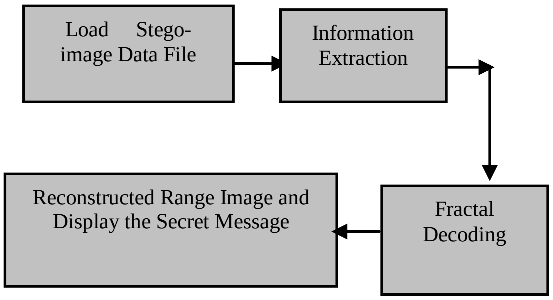

Load Stego-image Data File
Information Extraction
Fractal Decoding
Reconstructed Range Image and Display the Secret Message

Figure (4): Structure of extraction unit

2.2.6 Information Extraction

After the data file (stego-image) has been loaded, the process of reconstructing SecData is applied to extract the array of embedded secret characters, which have been stored in the (IFS) coefficients (s, o) in a reverse way. This stage implies the following steps:

1. Extract the two digit of the fraction part of the coefficient (s and o) with keeping the integer part.

$$V = IFS(I).scl - Fix\ (IFS(I).scl)$$

$$Vs = v \times 1oo$$

2. Convert the extracted data to byte

$$SecData(I) = CByte((\ Vs - Fix(Vs)) \times 1000$$

3. Convert the bytes to string representation

$$recSecData = recSecData\ \&\ CStr(Chr(SecData(I)))$$

4. Display the secret message.

3. System Implementation

The goal of this system is encoding and embedding or hiding information (text, numbers, symbols or equations) in a cover-image (BMP format) after compressing the image to produce the stego-image as a data file. System implementation accepts six inputs in the embedding stage**:**

1. Input Text to be Encoding and Hiding
2. Encoding the text by using new IFS cryptography approach
3. Loading the cover image (BMP. format) as the input file
4. Input control parameters
5. Quad-tree partitioning the colour component
6. Domain generating
7. Inputting the secret message for embedding
8. Fractal encoding which include embedding.

System implementation accepts one input in the extracting stage:

1. Input the data file which contains the secret message beside image data array.
2. Extracting the secret message and reconstructing the image in the same time.

3.1 System Requirements

The Microsoft Window XP has been used as an operation system and Visual Basic (VB6) as a programming language.

3.2 System Steps

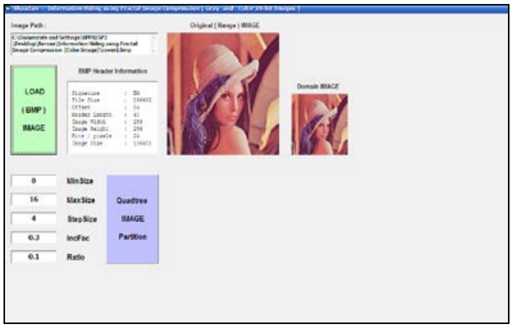
The proposed system steps:-

- **Input Secret Message**
- **Encoding Stage by using New IFS Cryptograph method.**
- **Partitioning the Cover Image**

After the cover-image has been chosen, control parameter will be entered to perform quad-tree partition for each colour component (R component, G component, B component) separately.

- **Generating the Domain Image**

Generating the domain image and domain pool is next to the partition step, the domain size taken is quarter the image size with overlapped blocks.

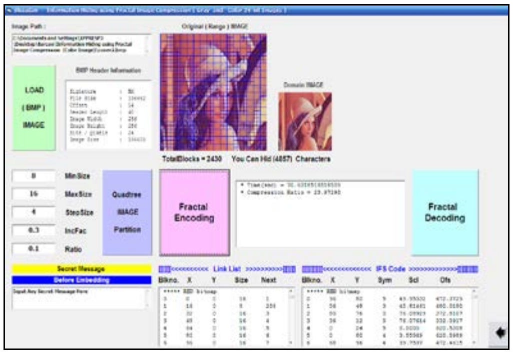


MinSize 8
MaxSize 16
StepSize 4
IncFac 0.3
Ratio 0.1

Figure (5): Domain image generating

- **Fractal Encoding**

Searching for similarity is performed between the range and the domain blocks and the information is stored in an index, then the image (cover-image) information is stored as a structure array of data.



TotalBlocks = 2430 You Can Hid (4857) Characters

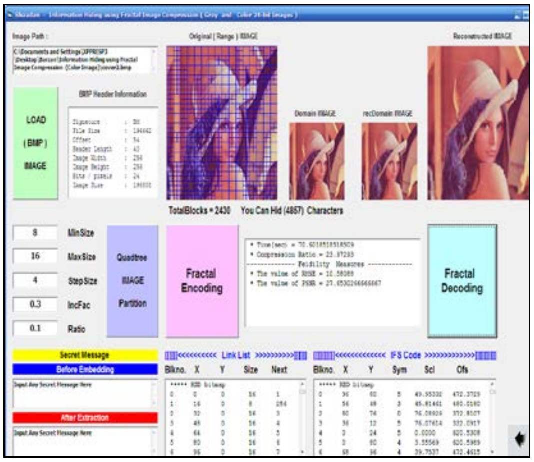
MinSize 8
MaxSize 16
StepSize 4
IncFac 0.3
Ratio 0.1

Input Any Secret Message Here

Figure (6): Fractal encoded data

- **Fractal Decoding**

The received data is a collection of data that represent the image with the secret message. The receiver will extract the embedded information (secret message) and then reconstruct the cover image.



TotalBlocks = 2430 You Can Hid (4857) Characters

MinSize 8
MaxSize 16
StepSize 4
IncFac 0.3
Ratio 0.1

```
• Time(sec) = 70.6018518518509
• Compression Ratio = 23.37293
------------- Feidility Measures -------------
• The value of MSE = 10.58089
• The value of PSNR = 27.8530266666667
```

Input Any Secret Message Here

```
Blkno.  X    Y    Size  Next
***** RED bitmap
0       0    0    16    1
1       16   0    8     256
2       32   0    16    3
3       48   0    16    4
4       64   0    16    5
5       80   0    16    6
6       96   0    16    7
```

```
Blkno.  X    Y    Sym  Scl       Ofs
***** RED bitmap
0       36   60   5    49.95332  472.3729
1       56   48   3    45.81461  480.0180
2       60   74   0    76.08929  372.8507
3       36   12   5    76.07614  332.0917
4       2    24   5    0.0000    620.5308
5       2    60   4    3.55569   620.5969
6       68   36   4    39.7537   472.4615
```

Figure (7): Secret message extraction and cover image reconstruction

3.3 The Effect of Hiding Secret Message

The main objective of the proposed hiding scheme is to embed a secret message with a huge number of characters as possible with different languages and numbers without degrading the quality of the reconstructed cover image. So to evaluate the effect of the secret message embedding on the cover image, a set of tests is applied. **Table (3)** show the result of hiding different message process.

Table(3) Hiding effect on Lena image

Max=8, Min=4, StepSize=4, R_a = 0.1, I_f =0.3, PSNR1= Stegoimage fidelity,PSNR2= After extraction fidelity, PSNR3= Without embedding					
N0.char	**Type**	**Time(sec)**	**PSNR1**	**PSNR2**	**PSNR3**
9653	**English-text**	**100.78**	**31.056**	**31.056**	**31.064**
10283	**Arabic-text**	**101.78**	**31.694**	**31.055**	**31.064**
11390	**Mixed**	**102.94**	**31.0644**	**31.064**	**31.064**

Table (4) Testing the effect hiding on Baboon image

Max=8, Min=4, StepSize=4, *Ar* = 0.1, *If*=0.3, PSNR1= Stegoimage fidelity, PSNR2= After extraction fidelity, PSNR3= Without embedding					
N0.char	**Type**	**Time(sec)**	**PSNR1**	**PSNR2**	**PSNR3**
9653	**English-text**	**116.54**	**23.246**	**23.246**	**23.246**
10248	**Arabic-text**	**117.85**	**23.246**	**23.246**	**23.246**
19950	**Mixed**	**118.74**	**23.246**	**23.246**	**23.246**

4.0 Conclusions

1. In this paper we have presented a new method to design a cryptographic system utilizing fractal theories. This approach employs two level methods, the firstly by using new approach fractal cryptography (encoding and decoding), and the secondly by hiding the encoding text by using proposed fractal image compression , which make the decoding more difficult, by embedding the attractor in a colored image using the LSB; and sending it to the recipient to decode the colored image and applying the key agreement to get back the message characters, by the collage method. This way to hide information is very useful cause even if the third party (intruders), recognized that there is a difference in the received image, wont figure what its , whether a lose in the information or just a rubbish data.

2. For embedding the secret massage, a new method has proposed and in this new method a huge number of characters can be embedded, beside the characters or the secret message there may not necessary be a text, it may consist of equation or numbers with another language.

3. The proposed system does not affect the image quality; we can say it is not noticeable for human eyes. To prove this we show the cover-image and the stego –image to a team of 15 persons to take their opinion if there is any difference between the stego-image and the cover-image and their answer that there is no difference between both images.

References

[1] Alia M., Samsudin, A., “**A New Approach to public-key cryptosystem based on Mandelbrot and Julia**”, Ph.D. Thesis Universiti Sains Malaysia, 2008.

[2] Kocarev, L., Sterjev, M., Fekete, A. and Vattay, G. “**Public-key encryption with chaos**”. *Chaos.* 2004, **14**(4):1078-82.

[3] Kumar, S. “ **Public key cryptography system using Mandelbrot sets**”, *Military Communications Conference, 2006.* MILCOM 2006. IEEE. 23-25 Oct.

[4] Gulati, K and Gadre**,** V.M. **“ Information Hiding using Fractal Encoding”**. Dissertation for the degree of Master of Technology. School of information Technology. Indian Institute of Technology Bombay. Mumbai, 2003.

[5] Fadhil Salman Abed, Nada Abdul Aziz Mustafa**, “ A proposed Technique for Information Hiding Based on DCT”,** International Journal of Advancements in Computing Technology Volume 2, Number 5, December 2010.

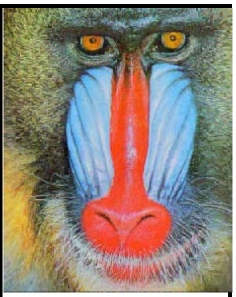
[6] Y. Fisher, “**Fractal Image Compression: Theory and Application**”, Springer-Verlag, New York, NY, USA, 1995.

[7] Fadhil Salman Abed,”**Adaptive Fractal Image Compression**”, Ph.d Thesis, Al-Rasheed College of Engineering and Science, University of Technology, 2004.

[8] Sua’d Kakil Ahmad, “**Image in Image Hiding System Using Iterated Function System (IFS)**”, Msc Thises, University of Sulaimani, 2009.

[9] Manoj Kumar Meena, Shiv Kumar, Neetesh Gupta**,”Image Steganography tool using Adaptive Encoding Approach to Maximize Image Hiding Capacity**”, International Journal of Soft Computing and Engineering (IJSCE) , ISSN: 2231-2307, Volume-1, Issue-2, May 2011.

Fadhil Salman Abed is a lecturer at the Depratement of Computer Sciences, Technical Institute of Kalar.. He received the B.Sc. degree in Mathematic from the University of Basra, Iraq, in 1987. He obtained his M.Sc. in Applied Mathematic(Computer Security) from University of Technology in 1997 and Ph.D. degree in Applied Mathematic(Fractal Image Compression) from University of Technology in 2004 . His research interests are in the field of Cryptography, Image Processing, Network secur has many research papers in Image Processir computer security.

	
Original Lena image256×256	**Original Baboon image256×256**

Table(6): The Some of Tests mages

Apendix

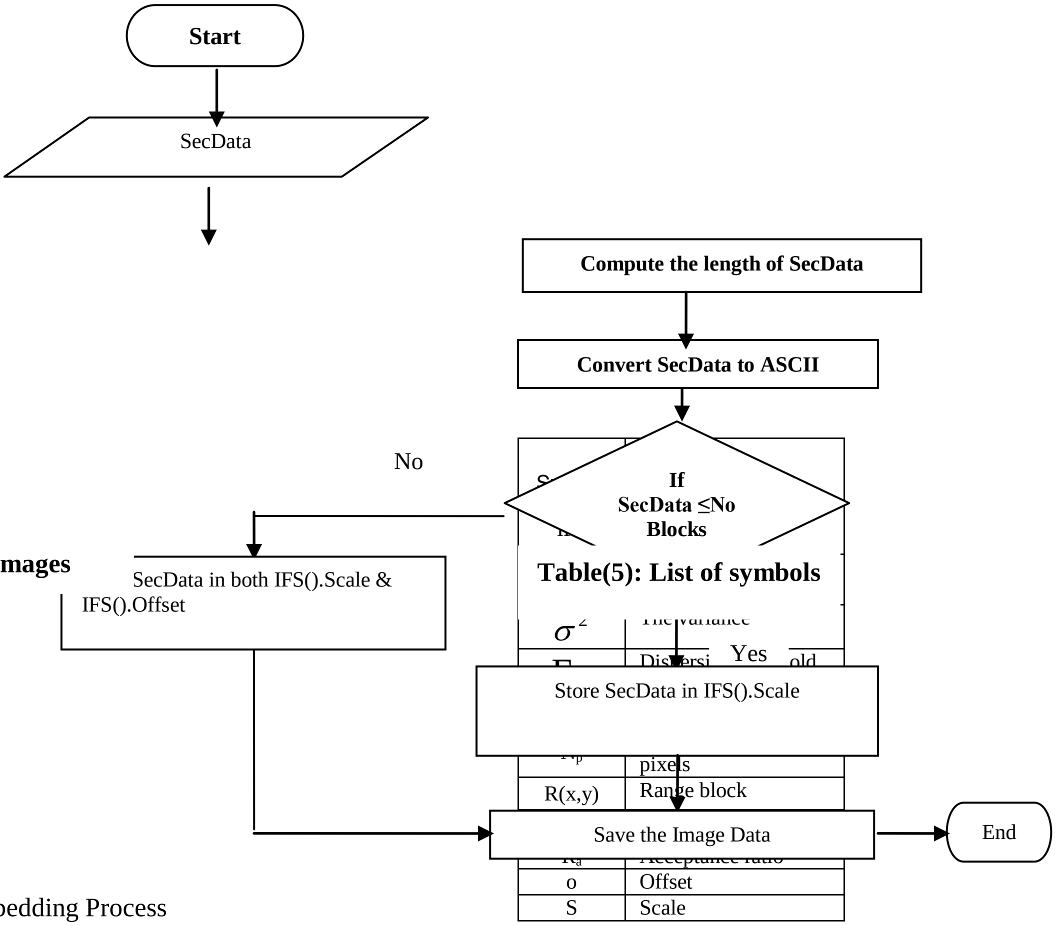

Start
SecData
Compute the length of SecData
Convert SecData to ASCII
If SecData ≤No Blocks
No
Yes
SecData in both IFS().Scale & IFS().Offset
Store SecData in IFS().Scale
Save the Image Data
End

Figure (8) Flowchart of Embedding Process

Table(5): List of symbols

σ^2	The variance
E	Dispersi old
N_p	pixels
R(x,y)	Range block
R_a	Acceptance ratio
o	Offset
S	Scale

Design and implementation of a platform for location-based services: a case study of GIS of archaeological and handicraft of Fez Medina

Abdesslam ESSAYAD

University Sidi Mohamed Ben Abdellah, Faculty of Siences Dhar El Mahraz Fez
BP 1796 Atlas Fez, Morocco

Abstract

The location-based services (LBS) allow users on the go with access to geographic data from micro-laptops, tablet PCs, personal digital assistants (PDA). These services refer to technologies GPRS2 UMTS3 and can use data on the spatial location of GPS devices in map. This paper presents a platform dedicated to the Medina of Fez, which will be an addition to the Moroccan strategy of development of handicrafts and tourism, called 'Vision 2015 Craft' [1] and 'Vision 2020 tourism'[2] which aims to raise Morocco among the 20 leading tourist destinations. The map used is of the Medina of Fez in SVG, with the language J2ME and J2EE technology.

Keywords : *Localisation Based Servic, GPS, J2EE, J2ME, SVG.*

1. INTRODUCTION

The location-based services(LBS) has great potential in identifying points of interest to the location on a map as desktop applications or mobile. The design and implementation of these services target geolocation in combination with geographic information useful to provide relevant content to users on the site.

The generalisation of the rich map geographic information in real time meets the needs of specific users [3].

This platform can provide the following services : information on archaeological sites and artisan of the Medina and other information as the query.

This platform will be divided into client and server side. The client side of this software platform is developed on the basis of the following : SVG and JSR179 JSR226 based mobile for mapping, GPS and J2ME. The server side is developed by XML, J2EE, and MySQL.

Fez is known worldwide for its long history, its historical heritage and crafts, culture, where she wowed millions of visitors in recent decades, tourism has continued to grow(almost 1 million visitors per year) has become an important economic factor. The extension of this type of technology to bring tourism benefits the local economy.

The Medina (old city) of Fez was founded in the eighth century, it was the first site of the country declared world heritage by UNESCO in 1981. Its architecture is particularly rich; archaeological area and crafts, bazaars, mosques, madrasas, tombs and palaces ....

This document is divided into four sections are organised as follows: the introduction, Section 2 gives a brief explanation of LBS technology and its implementation in the Medina. Section 3 announces the location technique. Section 4 describes the techniques used, the final section detailing the architecture and functions available in the system and finally the conclusion.

2. THE LOCATION BASED SERVICES (LBS)

The location of a mobile agent is one of the important issues of modern mobile communication system.

Location-based Services(LBS) are services to provide information stored in a database. This information can be created, compiled, selected, or filtered in the light of the current location of the mobile user. In each mobile currently the information delivery service has become indispensable. LBS allows us to find the geographical localisation of the mobile device is the GPS coordinates (latitude and longitude), and offer services based on this location information is the current place in the map.

You can now have a mobile device presents the map data where the user will always be the center of the map and the new standards of data vectors will improve and streamline the transfer of map data to a mobile user.

The development of mobile technology has had a significant impact on services and other human activities such as tourism-related activities. Understanding of market opportunities in tourism and the increased demand facing the mobile business can make the technique of a localisation phenomenon of the day.

Revenues from these services will grow about 25.5 billion dollars in 2008 [4]. The basic idea of these services is to locate the agent on a map and provide the information to choose according to his profile. When the agent is in a location that does not know, his need is to find a place to sleep,eat, also find an ATM to withdraw money. As the tourism sector is heterogeneous, the diversity of information services for mobile users is clearly a question of ease of use and meet all needs.

LBS technology is the intersection between three technologies : information technology and telecommunication(ICT), the mobile telecommunications system and geographic information system (GIS).

LBS adds structure to the real world for the fixed Internet and wireless networks with dynamic content related to the location [5].

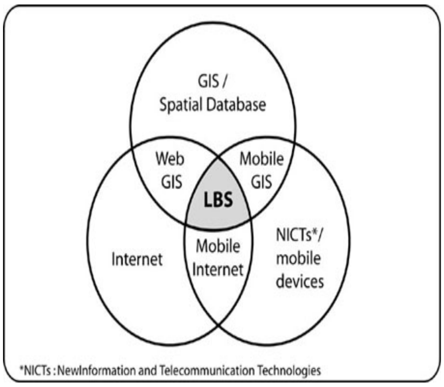


Fig. 1: LBS is crossroads of technology [6]

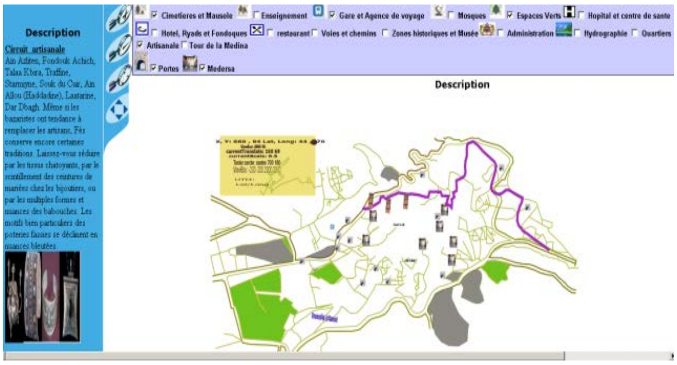
On the platform there are two different ways to access services LBS :

1) *The initiative of the platform:*

The user sends a request (a text) for information on services in areas near him, or craft and historic areas.

2) *The initiative of the user*

The user registers in advance to receive certain information each time it is close to a place of interest. He can receive the requested information on new items or promotions craft when it is near the place.



☑ Cimetieres et Mausole □ Enseignement ☑ Gare et Agence de voyage □ Mosques ☑ Espaces Verts □ Hopital et centre de sante
□ Hotel, Ryads et Fondoques □ restaurant □ Voies et chemins □ Zones historiques et Musée □ Administration □ Hydrographie □ Quartiers
☑ Artisanale □ Tour de la Medina
☑ Portes ☑ Medersa

Description

Circuit artisanale
Ain Azliten, Fondouk Achich, Talaa Kbira, Traffine, Stamyne, Souk du Cuir, Ain Allou (Haddadine), Lattarine, Dar Dbagh. Même si les bazaristes ont tendance à remplacer les artisans, Fès conserve encore certaines traditions. Laissez-vous séduire par les tissus chatoyants, par le scintillement des ceintures de mariées chez les bijoutiers, ou par les multiples formes et nuances des babouches. Les motifs bien particuliers des poteries fassies se déclinent en nuances bleutées.

Description

Fig. 2: Extract from the Medina map showing the artisanal circuit

2.1 SERVICES

2.1.1 *The service itinerary*

The application helps customers plan their travel, manage their time and reach their destination by describing the exact area.
Subscribers can request the road (text) of their trips.

2.1.2 *The proximity service*

The proximity services allow subscribers to search and/or to identify points in their neighbourhood public interest such as archaeological sites, museums, monuments, handicrafts, car parks, and other public services, etc ...

2.1.3 *Rescue service*

This service enables a subscriber in trouble to call in an emergency to a service that can locate and to provide the necessary assistance. This assistance may cover the following needs: auto repair, medical emergency and police assistance.

2.1.4 *Information on archaeological sites and artisan*

Information on archaeological sites and artisan are also stored in the database server. The requested information is a request from the terminal to the server via the HTTP protocol.

2.1.5 *Hotel Information*

The user can enter the name of the hotel, via a form that will be sent to the server, the server stores it and returns a description sheet with relevant information, including text and images.

2.1.6 *Browsing*

After downloading the map, the user can move the map in the lower left and right, the map data is stored and will be displayed as sub-layers, the user can automatically show or hide certain layers.

2.1.7 *Related work*

Today there are hundreds of sites and mobile applications that offer services based on the current position of the users. These include "Google Latitude" which allows you to specify which one is in real time geolocate his friends by sending them an invitation in advance. "Twitter" and soon "Facebook" also allow users to share their location. Finally,"Foursquare" or French "Plyce" allow people to identify, from their mobile phone, his friends who are nearby. Users can also recommend and advise on their favorite places such as restaurants or coffee.

The work of this research emphasizes the approach that provides services dominant in a definite place.

3. GEOLOCATION

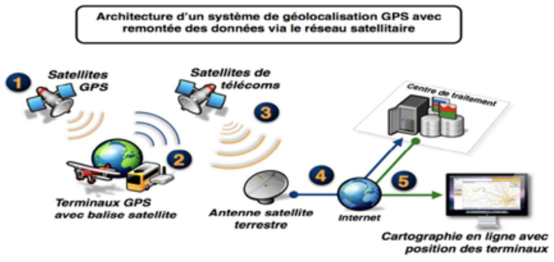
3. 1 POSITIONING TECHNOLOGIES

There are several methods to locate a mobile, which are based on the transmission of certain signals and their reception at the other end. Positioning technology used is selected according to need (e.g. response time) applications. Among the various existing positioning technologies, we have : GPS Global Positioning System: uses a range of satellites to locate the user. This raw information is processed through the terminal or sent over the network to be processed to recover the position. The accuracy of the position varies between 5 and 40m clear sky.

The current position must be to the extent possible, meaningful information is based on the principle of reverse geocoding of translating the GPS data (longitude, latitude) to address human-readable. This procedure for reverse geocoding is limited initially to translate the position based on the available database.

With Java ME a class can receive the GPS coordinates. In fact from the start of a Midlet application contains three text fields to receive the GPS coordinates (Longitude, Latitude and altitude) automatically.

The J2ME API[7] for the JSR179 specification defines a group of options, javax.microedition.location, which allows developers to write location-based applications and services for devices of limited resources such as mobile phones.


Architecture d'un système de géolocalisation GPS avec remontée des données via le réseau satellitaire
1
Satellites GPS
Satellites de télécoms
3
Centre de traitement
2
Terminaux GPS avec balise satellite
Antenne satellite terrestre
4
Internet
5
Cartographie en ligne avec position des terminaux

Courtesy of Novacom Services - Image under Cc-by-sa-3.0 license

Fig. 3: Architecture of a GPS location system with feedback of data via the satellite network.

3. 2 CONVERTING GPS COORDINATES I N SVG

* Convert sexagesimal degrees in decimal degrees.

The geographical coordinates are often given in sexagesimal degrees, ie in degrees,minutes and seconds. However, computers prefer the decimal system and it is necessary to convert sexagesimal degrees in decimal degrees.

Example. Is a latitude of 45 ° 53 '36 " (45 degrees, 53 minutes and 36 seconds). Expressed in decimal degrees, the latitude is equal to : latitude = 45 + (53 / 60) + (36 / 3600) = 45.89

General formulation: latitude (decimal degrees) = degrees + (minutes / 60) + (seconds/ 3600)

* Create a scale of data: That is to define the range of min and max GPS that was, andput them in the SVG using a function y=a *x + b. In the database, there is a minimum longitude L_min and maximum longitude L_max. When the minimum and maximum longitudes are known, ranging from -5.004101 to -4.952431 for about Medina and if the map is 500px wide, so just find the values of a and b from the two following equations.

0 = a * (-5.004101) + b; 500 = a * (-4.952431) + b
This gives a= 9676.79 and b=48423.66. To place items in the database, it is necessary to apply the function :

x= 9676.79*longitude +48423.66.

For example, if you have a point with longitude -4.970026, will be placed in 330px on the SVG map.

This small system of equations is sufficient to cover all longitudes in the right place on the map SVG.

The same for the latitude, you can have two functions to link latitude and longitude with height and width of the SVG map.

3. 3 DISTANCE BETWEEN TWO POINTS: ROUTE SIMULATION

1) *formula of Haversine*

The calculation of the estimated distance between the locations of the points is important to manage time and organise itinerary. Haversine formula is preferable to be used in GIS applications, it assumes a spherical earth and does not include the effects ellipsoidal. To calculate the distance between two terrestrial coordinates as shown in Figure 4, the following algorithm is used:

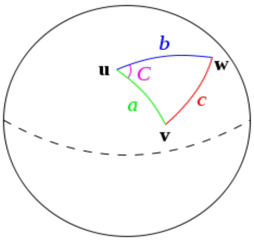


Fig. 4: Spherical Triangle to illustrate the law of Haversine[10]

The expression of Haversine formula is:

$$\text{haversin}\left(\frac{d}{R}\right) = \text{haversin}(\varphi_2 - \varphi_1) + \cos(\varphi_1)\cos(\varphi_2)\,\text{haversin}(\Delta\lambda).$$

Haversin is Haversine function Haversin (θ) = sin 2 (θ / 2) = (1-cos (θ)) / 2

d is the distance between two points (along a great circle of the sphere, see spherical distance).

R is the radius of the sphere (R = 637 100: the radius of the earth in meters).

φ1 is the latitude of point 1,
φ2 is the latitude of point 2
Δ λ is the separation of longitude,

On the left side of the equals sign, the argument of the function Haversine is in radians.

In degrees, Haversin (d / R) in the formula would become Haversin (180 ° d / π R).

We can then solve for d either by simply applying the inverse Haversine (if available) or by using the arcsinus (inverse sinus):

$$d = R\,\text{haversin}^{-1}(h) = 2R\arcsin\left(\sqrt{h}\right).$$

h is Haversin (d / R)

The implementation of this formula in a MIDlet class is as follows:

Δlatitude = φ1 - φ2

Δ λ = long1 – long2

a = sin2(Δlatitude / 2) +

cos(φ1) * cos(φ2) * sin2(Δ λ / 2)

c = 2 * atan2(√a, √(1 - a))

d = R * c

2) shortest path

The Dijkstra algorithm finds use in calculating the exact itinerary, However the Haversine formula calculates the estimated distance no longer interested nodes, it is a distance as the crow flies. As against this algorithm requires the need of information stored in the database with the weight of the arcs, the distance (the shortest route), the estimated time (for the fastest route), the most cost etc ...

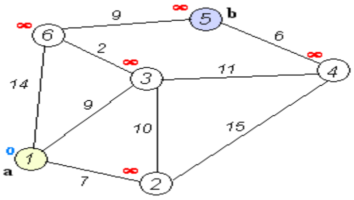


Fig. 5: An example of the progress of Dijkstra's algorithm

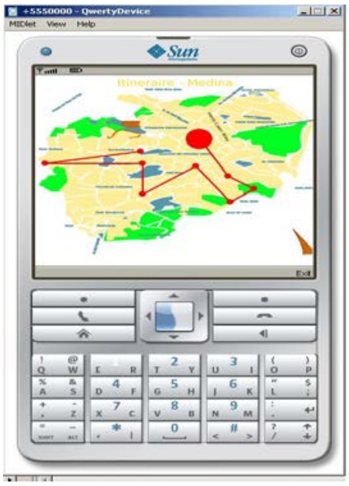


Fig. 6: Extract of a route traveled

Assume that L (x) denotes the length between the node 1 to node 5,a Java class dedicated to implement the algorithm returns the shortest path and distance.

The platform provides the shortest route between two places after the entry of the place name, the server can calculate the shortest path and return to the client the exact path. The coordinate information will be added to file SVG to redraw the map and display the path on the terminal screen.

3. 4 IMPLEMENTATION

The architecture can be divided into two main parts: the local client and remote server.

The client manages data visualisation locally on a Web page for the PC or using a MIDlet for mobile phones. The user requests and communication with the remote data server can perform the loading of the SVG file that is the map with the required choice.

The platform includes mobile equipment like PDA or PC and the remote Web server. The Web server forwards the client execution environment for data visualisation. The server responds to user queries by running them itself or by redirecting them to the databases.

To extract data from the map SVG global XSLT processing is done on the server by the Apache Xalan processor, which uses Xerces for parsing XML documents. Servlets are simply receiving the request, invoke the Xalan processor to process and return the answer. In this architecture, the data received by the client are in HTML and SVG.

They are usable with any browser plug-in Adobe SVG Viewer.

Also a Midlet dedicated to mobile devices is responsible to download an excerpt from the SVG global map from the server after authentication.

Another solution is Google Static Maps API is an easy way to provide a map when the user does not have Javascript available. It is not as powerful as the full Google Maps API, but it can provide a base map containing both markers and paths. The basic concept is to generate the image by adding parameters to the query string of the URL. The link of the Google Static Map must be in the form below to respond to the request of the API:

“http/maps.google.com/maps/api/staticmap?parameters ”

The platform helps to customize these settings.

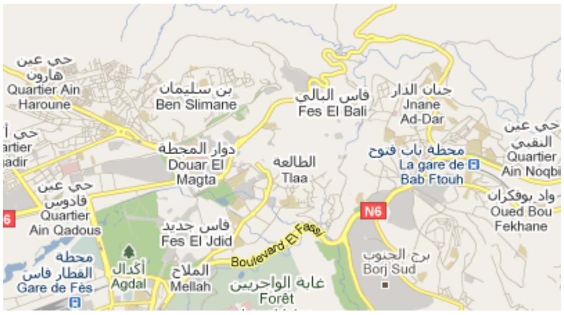


Fig. 7: Extract image Google Static Map

The Fig.7 shows the creation of a picture of a Google Map centered on the Medina of Fez. The url of the image is :

http://maps.google.com/maps/api/staticmap?center=Fez,Ma&zoom=13&size=900x900&sensor=false.

Other Google services implementation of the system is the geocoding to convert geographic coordinates to address.
Geocoding API supports reverse geocoding directly using the following URL :

http://maps.googleapis.com/maps/api/geocode/json?latlng=x,y&sensor=true_or_false
x is the latitude and y is longitude. The result will be a JSON(JavaScript Object Notation) which later will be analyzed by the server to return a specific address of the current location.

Reverse geocoding is to determine the address of a point on a map. This technique works well in urban areas and in countries well documented (geographically).

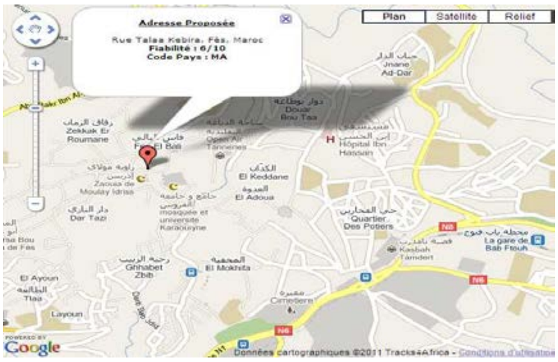

Adresse Proposée
Rue Talaa Kebira, Fès, Maroc
Fiabilité : 6/10
Code Pays : MA

Fig. 8: reverse geocoding with Google

4. TECHNICAL

4.1 CLIENT-SERVER MODEL

The choice of J2EE technology enables the development of applications that can distribute and run on a set of platforms.

In this work, the term customer refers to all hardware resources(laptop, PDA, PC ...) and software(MIDlet, Navigator) used by the mobile user to access system services.

The term server is used to name the computer on which the resources are centralised (SVG Map of Medina, Mysql Database,MVC Model-
based JSF ....), That means that data and applications, which accessed by the client.

The user (client) and uses a local application or MIDlet or browser,connected to the network via a wireless (WiFi, 3G, GPRS ...)

4.2 SVG

Scalable Vector Graphics (SVG) is a format that is known to have great potential to play an important role in the visualisation of geospatial data. SVG is a standard language developed by the W3C (World Wide Web Consortium) for describing two-dimensional graphics in XML[9].

SVG is a language for describing two-dimensional vector may havevector/raster graphics in XML.

The W3C publishes two recommendations: SVG1.1 and SVGMobile.

This open vector graphics format extends the multimedia capabilities of mobile phones and handheld computers.

4.3 J2ME

In J2ME, the Java runtime environment is suitable for machines whose capabilities are limited.

In this platform this language creates a class Midlet dedicated to mobile devices at the end to connect to the server and have required result.

HttpConnectiom class provides the communication link between the client and the server from the programming language and supportsJ2ME JDBC connection to access the server database, the behavior is one that combines HttpConnection InputStream and OutputStream.

4.4 XSLT LANGUAGE

XSLT (**eXtensible Stylesheet Language Transformations**) makes the transformation to create a new XML or HTML document from an XML document. It is mainly used to create different views of a document for presentation to a specific user, but it can be used to make treatment more complex. XSLT and XSL-FO is set XSL (**eXtensible Stylesheet Language**), the language of style sheets for XML.

The transformation made by XSLT file from a global file SVG to file SVG reduced, because of its implementation in a mobile device.

4.5 GPS

GPS (Global Positioning System) is now the preferred tool to position with great accuracy and also serves as a time reference.

The system Navstar GPS (Navigation System by Timing And Ranging-Global Positioning System), commonly called GPS, is a positioning system using satellites. This system is global because it can be positioned at any time and anywhere in the world and its space environment.

Note that the precision HF is only accessible from May 2, 2000.

5. ARCHITECTURE OF THE PLATFORM

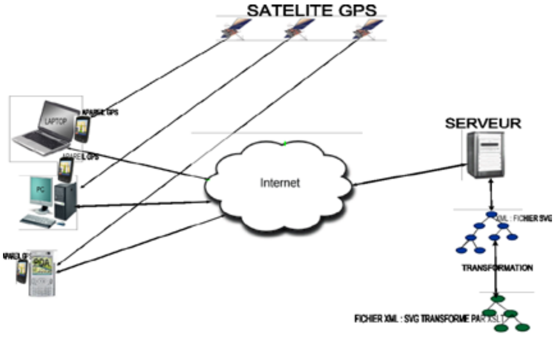


Fig. 9: Platform architecture

The basic concept for the implementation of this work along two axes. The first is to use the new XML technologies in a graphic in a Web environment combining traditional formats and raster images and the other is to prove that this architecture is easily implemented and gives the end-user effective results.

Server: It is in charge of analysing the request and resolve data transferred from the wireless network, and to seek information from the database, and generate a good response.

Database: It is used to store archaeological and craft of the card and paths, user information and profile, and other geographic information useful for viewing maps with SVG. The data stored in the system consist of two parts:

One is the location information, related to the user coordinates that will be instantly stored in a table at the end of his route and find areas that interested him.

The other is submitted to the user from the database. The data are generated after the user made its request. LBS accepts user data and transfers them to the server, the server executes the query and the answer to the customer from the database.

The server can generate the type of map chosen from the start of Midlet, the extraction is done from a global XML file, this file contains data areas, routes to the destination of the zones and information, hotels etc. All data exchanges between the system and the terminal are based on XML on interaction with a MySQL database.

6. CONCLUSIONS

This platform provides customers with LBS services to mobile customised to its current location. Localisation services will be more important factors in the future. A platform for LBS concept is a development based on the convergence of several technologies including SVG Mobile, J2ME, J2EE, etc.

Localisation performance can be enhanced by the assistance of the provider network and updating of data should also be taken into account in practical application. In this article we have introduced location-based services used in the field of mobile telephony, these services are used to route the archaeological information and craft at all times and make available to the user based on its location.

ACKNOWLEDGEMENT

I thank my supervisor Dr. Rachid BENSLIMANE who gaves me guidance during this research. His advices were very helpful.

References

[1]. Plan de Développement Régional de l'Artisanat Région Fès – Boulemane 2007-2011

[2]. La vision stratégique de développement touristique (vision 2020)

[3]. Kupper, A. (2005). Location-Based Services Fundamentals and Operation. Wiley.

[4]. U.S. Wireless Business Location-Based Services 2006-2010 Forecast IDC - 9/29/2006 - 21 Pages - ID: IDC1375802

[5]. Jochen Schiller, Agnès Voisard, " Location-Based Services", *Morgan Kaufmann*, April 30, 2004

[6]. Steiniger, Stefan, Moritz Neun and Alistair Edwardes, Foundations of Location Based Services

[7]. JSR 179: J2ME Location API. *http://jcp.org/en/jsr/detail?id=179*

[8]. J. David Eisenberg. *SVG Essentials*. O'Reilly, Sebastopol California, USA. 2002.

[9]. Jon Ferraiol editor. Scalable Vector Graphics (SVG) 1.1 Specification. Available at http://www.w3.org/TR/2003/REC-SVG11-20030114/. 2003.

[10]. http://upload.wikimedia.org/wikipedia/commons/3/38/Law-of-haversines.svg

He received his license option electronics, university Sidi Mohamed Ben Abdellah in 1994 and Graduate Diploma of specialized in 2007 in the same university. He is currently chief of the Office Computers to the delegation of the Ministry of Education of El Hajeb after having had his diploma professional analyst, is a member of three national and local associations.

A Methodological Review for the Analysis of Divide & Conquer Based Sorting/ Searching Algorithms

Mr. Deepak Abhayankar[1] and Mrs. Maya Ingle[2]

**[1]School of Computer Science, Devi Ahilya University
Indore, M.P. 452017, India**

**[2]School of Computer Science, Devi Ahilya University ,
Indore, M.P. 452017, India**

Abstract

This paper develops a practical methodology for the analysis of sorting/searching algorithms. To achieve this objective an analytical study of Quicksort and searching problem was undertaken. This work explains that asymptotic analysis can be misleading if applied slovenly. The study provides a fresh insight into the working of Quicksort and Binary search. Also this presents an exact analysis of Quicksort. Our study finds that asymptotic analysis is a sort of approximation and may hide many useful facts. It was shown that infinite inefficient algorithms can easily be classified with a few efficient algorithms using asymptotic approach.

1. Introduction

There have been abundant computer applications which need sorting/searching as a key component. Since SQL operations use it as an internal database subroutine, all database applications gain advantage of an efficient sorting/searching algorithm. Also sorting/searching is a must for some rudimentary database operations like a creation of indices and binary searches. Sorting is functional in operations like finding closest pair, determining an element's uniqueness, finding k^{th} largest element, and identifying membership. Many practical applications in computational geometry need sorting. For instance sorting is used to find the convex hull in computational geometry [10]. Applications that need sorting/searching include supply chain management, bioinformatics and computer graphics. Since sorting/searching problem has a lot of importance in real world, hence it will be fruitful to evolve a practical framework or methodology for analysis of sorting algorithms.

This paper develops an intuitive framework or methodology for the analysis of sorting/searching algorithms. To achieve this objective an analytical study of Quicksort and searching problem was carried out. This effort explains that asymptotic analysis can be misleading if applied carelessly. This study provides a fresh insight into the working of Quicksort and Binary search. Also this study presents an exact analysis of Quicksort. Although there already exist a few average case analyses, majority of the attempts finish up as asymptotic analysis. Our study finds that asymptotic analysis is a sort of approximation and may hide many useful facts such as large constant factors which make any algorithm insane for practical purposes. It was shown that infinite inefficient algorithms can easily be classified with a few efficient algorithms using asymptotic approach.

2. Searching an Analytic Study

It is not difficult to design a set of binary search like divide and conquer searching algorithms which lead to following recurrence.

$T(n) = c + T(nk/k+1)$

$T(1) = d$

Master theorem suggests the solution of the recurrence relation is $T(n) = O(\log n)$.

For $k = 1$ we will have recurrence relation for binary search. For $k = 2$ one gets ternary search. One of the observations of this study is that for $k > 1$ we can produce a sequence of increasingly inefficient algorithms by incrementing the value of 1. But asymptotic analysis puts all the algorithms in the same set. In fact all algorithms can flaunt logarithmic time complexity. It is the constant factor that differs. The key conclusion is that constant factor matters and one cannot blindly trust the asymptotic order. The algorithm designer has to examine the situation thoughtfully. Too high a constant factor will render an algorithm useless with certainty. This endeavor finds that an exact analysis may provide better insight than what asymptotic analysis may offer.

3. Probabilistic Analysis of Quicksort

3.1 Review of Probabilistic Analysis

In probability theory, a probabilistic arrangement is defined by a sample space S and a probability measure p. The points of the sample space are the possible result of the experiment and are called elementary events. An event is a subset of the sample space. For instance, one event we may care about is the event that the first die comes up 1. Another is the event that the two dice sum to 7. The probability of an event is just the sum of the probabilities of the elementary events contained inside it [9].

A random variable is a function from elementary events to integers or reals. For instance, another way we can talk formally about these dice is to define the random variable Y1 representing the result of the first die, Y2 representing the result of the second die, and Y = Y1 + Y2 representing the sum of the two. We could then ask: what is the probability that Y = 6? [9].

One property of a random variable we often care about is its expectation. For a discrete random variable Y over sample space S, the expected value of Y is: E[Y] = Pr(e1) Y[e1] + Pr(e2) Y[e2] +..............Pr(en) X[en] for all e∈S. An important fact about expected values is Linearity of Expectation: for any two random variables U and V, E[U+V] = E[U] + E[V]. This fact is incredibly important for analysis of algorithms because it allows us to analyze a complicated random variable by writing it as a sum of simple random variables and then separately analyzing these simple RVs[9].

3.2 Probabilistic Analysis of Quicksort with Accurate Results

Theorem 1The expected number of comparisons made by randomized Quicksort on an array of size n is Hn(2n+2) – 4n, where Hn = (1+ (½) +(1/3) +.............(1/n)).

Let us consider one of the random variables is Y_{ij}'s for i < j. Denote the i^th smallest element in the array by e_i and the jth smallest element by e_j. If the pivot we choose is between e_i and e_j then these two end up in different buckets and machine will never compare them to each other. If the pivot we choose is either e_i or ej then Computer does compare them. If the pivot is less than e_i or greater than e_j then both e_i and e_j end up in the same bucket and we have to pick another pivot. So, one can think of this like a dart game: we throw a dart at random into the array: if we hit e_i or e_j then Y_{ij} becomes 1, if we hit between e_i and e_j then Y_{ij} becomes 0, and otherwise we throw another dart. At each step, the probability that Y_{ij} = 1 conditioned on the event that the game ends in that step is exactly 2/(j − i + 1). Therefore, overall, the probability that Y_{ij} = 1 is 2/(j − i + 1).

$$Y = \sum_{i=0}^{n-2} \sum_{j=i+1}^{n-1} Y_{ij}$$

$$E[Y] = \sum_{i=0}^{n-2} \sum_{j=i+1}^{n-1} E[Y_{ij}]$$

$$E[Y] = \sum_{i=0}^{n-2} \sum_{j=i+1}^{n-1} \frac{2}{j-i+1}$$

Up to this stage we follow what the other researchers have already done [9], and from this point we move in the direction of exact value rather than a crude upper bound.

$$E[Y] = \left(\frac{2}{2}\right)(n-1) + \left(\frac{2}{3}\right)(n-2) + \left(\frac{2}{4}\right)(n-3) + \cdots + \left(\frac{2}{n-1}\right)(2) + \left(\frac{2}{n}\right)(1)$$

$$E[Y] = \left(\frac{2}{2}\right)(n-1) + \left(\frac{2}{3}\right)(n-2) + \cdots + \left(\frac{2}{n-1}\right)(n-(n-2)) + \left(\frac{2}{n}\right)(n-(n-1))$$

$$E[Y] = \left(\frac{2}{2}n - \frac{2}{2} \times 1\right) + \left(\frac{2}{3}n - \frac{2}{3} \times 2\right) + \cdots + \left(\frac{2}{n-1}n - \frac{2}{n-1} \times (n-2)\right) + \left(\frac{2}{n}n - \frac{2}{n} \times (n-1)\right)$$

$$E[Y] = 2n(H_n - 1) - 2(n - H_n)$$

$$E[Y] = H_n(2n+2) - 4n \quad \text{---------------- (Equation A)}$$

Equation A is one of the central contributions of the paper. Equation A gives the exact value of the expected number of comparisons performed by Quicksort. If a researcher is inclined towards asymptotic approach s/he can easily have it. For the researchers, who are inclined towards asymptotic approach and approximate results, E[Y]=O(nlogn). Because H_n is approximately log n, E[Y] becomes O(log n).

3.3 Alternative Analysis

This section is basically a byproduct of the overall study. It is a bit crude but effective technique for asymptotic analysis. Quicksort partition may divide the array into two partitions. One of the partitions may be empty. If there are two partitions then either both are of same size or one of them will be larger than the other one. We are interested in upper bound on the average case time. Size of the Non smaller partition may vary from (n-1) to (n-1)/2. Average size of Non Smaller partition was found to be approximately (3n/4). Along the same line if we estimate average size of Non large partition we get approximately (n/4). This leads to following recurrence relation.

$T(n) = T(n/4) + T(3n/4) + (n-1)$.

Application of Recursion tree approach recommends that solution is O(n log n).

4. Results and Conclusion

Evidence of the analysis of a set of divide and conquer search algorithms suggests that asymptotic analysis can easily mislead. Exact analysis is a better option than asymptotic approach. Asymptotic analysis may play a side role but it cannot replace exact analysis. If exact mathematical analysis is not feasible then only approximations and asymptotic can play the key role. References preferred to provide only the asymptotic analysis; this study seems to be unique to go beyond asymptotic analysis and to provide an exact analysis of Quicksort. Moreover this study produces one more alternative asymptotic analysis.

References

[1] D. E. Knuth, The Art of Computer Programming, Vol. 3, Pearson Education, 1998.

[2] C. A. R. Hoare, "Quicksort," Computer Journal5 (1) , 1962, pp. 10-15.

[3] S. Baase and A. Gelder, Computer Algorithms: Introduction to Design and Analysis, Addison-Wesley, 2000.

[4] J. L. Bentley, "Programming Pearls: how to sort," Communications of the ACM, Vol. Issue 4, 1986, pp. 287-ff.

[5] R. Sedgewick, "Implementing quicksort Programs," Communications of the ACM, Vol. 21, Issue10, 1978, pp. 847-857.

[6]T. H. Cormen, C. E. Leiserson, R. L. Rivest and C. Stein, Introduction to Algorithms, Second Edition. MIT Press and McGraw-Hill, 2001.

[7] G. S. Brodal, R. Fagerberg and G. Moruz, "On the adaptiveness of Quicksort," Journal of Experimental AlgorithmsACM, Vol. 12, Article 3.2, 2008.

[8] N. Wirth, Algorithms and Data Structures, © N. Wirth 1985 (Oberon version: August 2004)

[9]http://www.cs.cmu.edu/afs/cs/academic/class/15451-s10/www/Probabilisticanalysis, Randomized Quicksort-1-July-2011

[10] J. Chhugani, W. Macy, A. Baransi, A.D. Nguyen, M. Hagog, S. Kumar, V. W. Lee, Y. K. Chen, P. Dubey, “Efficient Implementation of Sorting on Multi-Core SIMD CPU Architecture , “ Journal Proceedings of the VLDB Endowment, Volume 1, Issue 2, August 2008.

Secure Authentication using Anti-Screenshot Virtual Keyboard

Ankit Parekh[1], Ajinkya Pawar[2] , Pratik Munot[3] and Piyush Mantri[4]

**[1,2,3] Department of Computer Engineering,
MIT College of Engineering,
Pune University, India**

**[4] Department of Computer Engineering,
Trinity College of Engineering and Research,
Pune University, India**

Abstract

With the development of electronic commerce, a lot of companies have established online trading platforms of their own such as e-tickets, online booking, online shopping, etc. Virtual Keyboard is used for authentication on such web based platform. However Virtual Keyboard still suffers from numerous other fallacies that an attacker can take advantage of. These include click based screenshot capturing and over the shoulder spoofing. To overcome these drawbacks, we have designed a virtual keyboard that is generated dynamically each time the user access the web site. Also after each click event of the user the arrangement of the keys of the virtual keyboard are shuffled. The position of the keys is hidden so that a user standing behind may not be able to see the pressed key. Our proposed approach makes the usage of virtual keyboard even more secure for users and makes it tougher for malware programs to capture authentication details.

Keywords: *Virtual Keyboard, Keylogging, Online Banking, Trojan Horse, Security, Efficiency*

1. Introduction

Today, the Internet has melted into our daily lives with more and more services being moved online. Besides reading news, searching for information and other risk free activities online, we are also accustomed to other risk related work such as paying using credit cards, online banking, etc. We enjoy the convenience but at the same time we are putting ourselves at risk. More and more Trojan Horses have developed which have been the huge trouble to security of e-commerce and makes businesses and customer suffer huge economic losses.

Virtual Keyboard is a mechanism used by many banks to solve the problem of their bank account and password being stolen by these Trojans. However the Virtual Keyboard is not foolproof and has its own fallacies. Hence, a new Anti-Screenshot Virtual Keyboard is proposed in the paper, which can protect bank accounts and passwords from stealing due to the screen capture of Trojan Horses.

2. Password Stealing Schemes

In this section various methods in which passwords are been stolen have been discussed. These methods are Shoulder Surfing, Phishing, Attacks using hardware and Attacks using Trojan Horses. These methods have been discussed in detail in Sections 2.1, 2.2, 2.3 and 2.4 respectively.

2.1 Shoulder Surfing

Shoulder Surfing is a well known method of stealing others passwords and other sensitive information by looking over victims shoulders while they are sitting in front of terminals. This attack is most likely to occur in insecure and crowded public environments such as Internet Café, shopping malls, etc. It is possible for an attacker to use hidden camera to record all keyboard actions of a user. Video of the users actions on the keyboard can be studied later to figure out users ID and Password.

2.2 Phishing

Phishers attempt to fraudently acquire sensitive information, such as passwords and credit card details, by disguising as a trustworthy person or business in an electronic communication. For example, a phisher may set up a fake website and send emails to potential victims and persuade them to access the fake website. This way, the phisher can get a clear text of victims password. Phishing attacks are proven to be effective.

2.3 Using Hardware

Hardware Keyloggers are used for keylogging by means of a hardware circuit attached somewhere between computer keyboards and the computer. All the keystroke activities are logged into the internal memory which can later be accessed. A Hardware keylogger has an advantage over software solution because it is not dependent on installation on target computer operating system, it will not interfere with any program running on the target machine and also cannot be detected by any software. However its physical presence can be detected.

Fig. 1 Hardware Keylogger

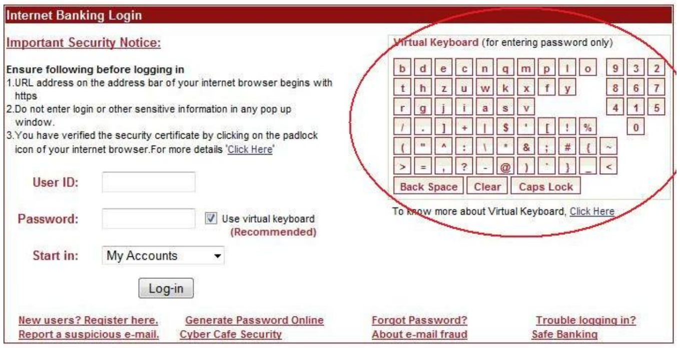
2.4 Using Trojan Horses

Trojan is a program that contains or installs malicious code. There are many such Trojans available online today. Trojans capture the keystrokes and store them somewhere in the machine and send them back to the adversary. Once a Trojan is activated, it provides the adversary with a string of characters that a user might enter online, consequently putting personal data and online account information at risk. They work in the background without the user coming to know about them. Chances of computer being affected by such malicious software is 70% even if the computer is up-to-date. A Trojan Horse typically contains two files: DLL file end EXE file. The DLL file does all the recording in some file in the computer while the EXE installs the DLL and triggers it to work. The file in which all the recording is done is mailed to the adversaries mail account.

3.Virtual Keyboard

Virtual Keyboard is a software technology that is used to mitigate the attack of password stealing Trojans. It is an on-screen keyboard which uses mouse to enter sensitive details such as an credit card pin number or password.

Fig 2 shows a virtual keyboard. The user has to use the mouse to click on the key on virtual keyboard he wants to press. The corresponding key will be typed into the selected textbox. Hence in this way using the virtual keyboard the use of traditional keyboard can be nullified. Thus sensitive and personal information can be protected.



Internet Banking Login

Important Security Notice:

Ensure following before logging in
1.URL address on the address bar of your internet browser begins with https
2.Do not enter login or other sensitive information in any pop up window.
3.You have verified the security certificate by clicking on the padlock icon of your internet browser.For more details 'Click Here'

User ID:

Password: ☑ Use virtual keyboard (Recommended)

Start in: My Accounts

Log-in

Virtual Keyboard (for entering password only)

b d e c n q m p l o 9 3 2
t h z u w k x f y 8 6 7
r g j i a s v 4 1 5
/ .] + | $ ' [! % 0
(" ^ : \ * & ; # { ~
> = , ? - @) ` } _ <
Back Space Clear Caps Lock

To know more about Virtual Keyboard, Click Here

New users? Register here. Generate Password Online Forgot Password? Trouble logging in?
Report a suspicious e-mail. Cyber Cafe Security About e-mail fraud Safe Banking

Fig. 2 Virtual Keyboard

4. Problems of Virtual Keyboard

The virtual keyboard has a number of fallacies that the attacker can take advantage of. There are advanced Trojans which take screenshots on the Mouse Click event. All these screenshots are uploaded to the hackers website or mail account. Hence in this way even a virtual keyboard is made susceptible to attacks. Also the virtual keyboard is susceptible to shoulder surfing.

5. Anti-Screenshot Virtual Keyboard (ASVK)

In order to overcome the fallacies of the virtual keyboard, such as susceptibility to screenshot capturing and shoulder surfing, the anti-screenshot virtual keyboard is proposed in this paper.

In the anti-screenshot virtual keyboard, when the mouse move to one key, all the keys on that particular row of the keyboard are changed to some special symbol like an asterisk(*) or a hash(#). Figure 3 shows the position of the anti-screenshot virtual keyboard when the mouse cursor moves on a particular key.

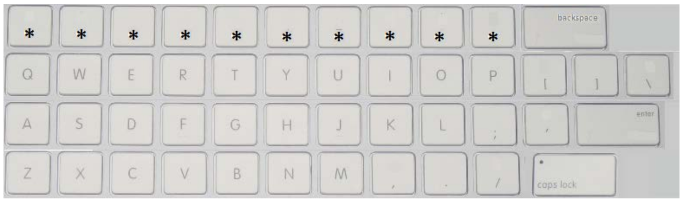

* * * * * * * * * *
backspace
Q W E R T Y U I O P [ ] \
A S D F G H J K L ; ' enter
Z X C V B N M , . / caps lock

Fig. 1 : Position of the virtual keyboard on mouse move event

When the user clicks a particular button all the keys on the virtual keyboard are changed to some special symbol, say asterisk(*). Hence even if the Trojan takes a screenshot on the mouse click event, all that will be captured is asterisk(*). Figure 4 shows the position of ant-screenshot virtual keyboard when a particular key is pressed.

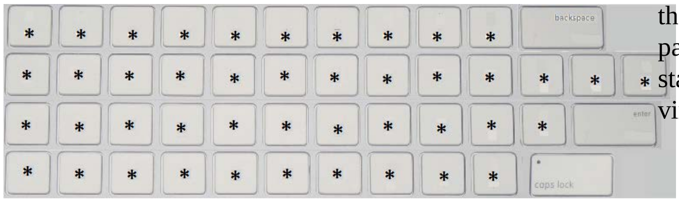


Fig 2: Position of the virtual keyboard on mouse click event

When the user releases the key, the keyboard is retained back to as shown in Figure 3.
However, in the above approach, if the Trojan Horse takes the screenshot of the virtual keyboard layout, then it is possible to identify the password. Hence in order to overcome this problem, real time refreshing can be done i.e when the user releases a key, instead of bringing the keyboard back to the original position, its keys can be randomized. However total randomization of the keyboard will make the user uncomfortable.
Hence in anti-screenshot virtual keyboard, the keyboard is divided into 15 areas as shown in Figure 5. The keyboard areas are as follows:

- 11 keyboard areas which consist of 3 keys.
- 3 keyboard areas which consist of 4 keys.
- 1 keyboard area which consists of 1 key.

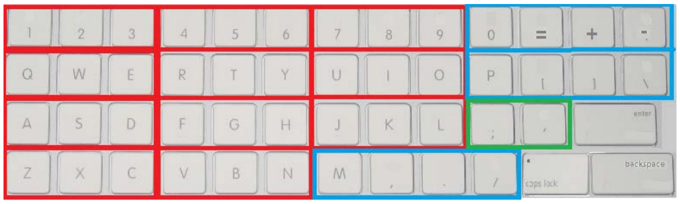

1 2 3 4 5 6 7 8 9 0 = + -
Q W E R T Y U I O P [ ] \
A S D F G H J K L ; ' enter
Z X C V B N M , . / caps lock backspace

Fig 3 : Division of Virtual Keyboard for randomization

Each time the page is loaded or a key is pressed by the mouse, the order of keys in each area is rearranged. Because of this randomization the Trojan cannot find the input content even though they have captured one or two pictures of virtual keyboard layout. Let us assume that the user has 8-bit password consisting of only characters, then the Trojan Horse will require two million attempts to get the password. The bank account is blocked after three unsuccessful attempts. Hence the level of security provided by the anti-screenshot virtual keyboard is very high.

6. Analysis of Anti-Screenshot Virtual Keyboard

The anti-screenshot virtual keyboard requires some delayed time of say 0.2 seconds after the mouse button is released to show the values of keys and they are refreshed by areas on the virtual keyboard. Thus it is obvious that efficiency of virtual keyboard is less than that of traditional keyboard. But it is worthy to make the account secure at the cost of efficiency. With the inputting of a 8–bit password the comparison of consumed time between standard traditional keyboard and the anti-screenshot virtual keyboard is as shown in Figure 6.

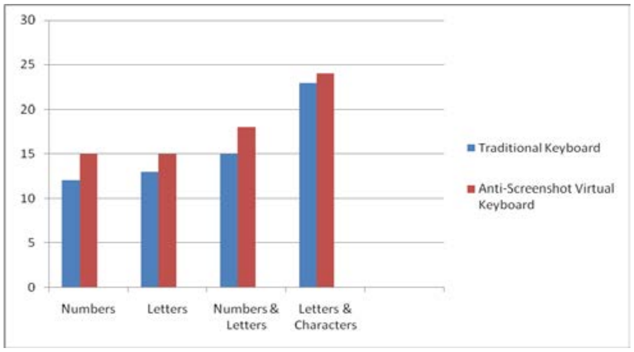


Fig. 4 : Efficiency Comparison

Results show that efficiency of virtual keyboards is 10.6% less than the traditional keyboards. When the customers input their accounts and passwords, the average time consumed is 20 seconds then 10.6% means 2 seconds slower than traditional keyboard. This 2 second delay will not make much difference to the customers because it is at the cost of security of account.

7. Advantages of Anti-Screenshot Virtual Keyboard

The anti-screenshot virtual keyboard is capable of solving the problem of screen capture by advanced Trojans. Also because of the change of caption of keys to some special symbols such as asterisk(*) or hash(#) its vulnerability to shoulder surfing is less as compared to the traditional virtual keyboards. Also the implementation of anti-screenshot virtual keyboard is simple. JavaScript can be used for its implementation.

8. Disadvantages of Anti-Screenshot Virtual Keyboard

The anti-screenshot virtual keyboard is less efficient as compared to the traditional keyboard. It is 10.6% less efficient as compared to the traditional keyboard. This makes the operation on anti-screenshot virtual keyboard slower. But it is worthy to make the account secure at the cost of efficiency.

9. Conclusion

The new design method of virtual keyboard presented in this paper to solve the problem of accounts of users are stolen because of Trojan Horses monitors the keyboards or captures the screen of internet payment is foolproof and improves the security and has an ability of becoming escort of online banking.

References

[1] Analysis of New Threats to Online Banking Authentication Schemes by Oscar Delgado, A. Fuster-Sabater and J.M.Sierra

[2] UCAM.CL.TR-731 ISSN 1476-2986 : A new approach to Online Banking by Matthew Johnson

[3] Matthew Pemble. Evolutionary trends in bank customer – targeted malware. Network Security, 2005(10):4–7, October 2005

[4] M.AlZomai, B.Al Fayyadh, A.J sang, and A.McCullagh .An experimental investigation of the usability of transaction authorization in online bank security systems. In Proceedings of the Australasian Information Security Conference(AISC'08), Wollongong, Australia, January 2008.

[5] Zhang Zhanjun, Xu Jialiang. Online virtual keyboard and intelligent input system, Tsinghua University Press, 1998.

[6] Video demonstrating a Trojan Attack against Caja Murica Bank of spain
http://www.hispasec.com/laboratorio/cajamurcia_en.swf

[7] Demo of Attack against Citi Bank India
www.tracingbug.com/index.php/articles/view/23.html

[8] Hyunjung Kim, Minjung Sohn, Seoktae Kim, Jinhee Pak, Woohun Lee , Springer. Button Keyboard: A Very Small Keyboard with Universal Usability for Wearable Computing. In Maria Cecília Calani Baranauskas, Philippe Palanque, Julio Abascal, Simone Diniz Junqueira Barbosa, eds. Human-Computer Interaction – INTERACT 2007, 11th IFIP TC 13 International Conference, Rio de Janeiro, Brazil, September 10-14, 2007, Proceedings, Part I. Lecture Notes in Computer Science 4662 Springer 2007. 343-346

X-ray view on a Class using Conceptual Analysis in Java Environment

Gulshan Kumar [1], Mritunjay Kumar Rai[2]

[1]Department of Computer Science and Engineering, Lovely Professional University

Phagwara, Punjab, India

[2]Department of Computer Science and Engineering, Lovely Professional University

Phagwara, Punjab, India[2]

Abstract

Modularity is one of the most important principles in software engineering and a necessity for every practical software. Since the design space of software is generally quite large, it is valuable to provide automatic means to help modularizing it. An automatic technique for software modularization is object- oriented concept analysis (OOCA). X-ray view of the class is one of the aspect of this Object oriented concept analysis. We shall use this concept in a java environment.

Keywords: *Concepts, views, dependencies, classes, methods and attributes, object-oriented properties*

1. Introduction

Concept Analysis (CA) is a branch of lattice theory that allows us to identify meaningful groupings of elements (referred to as objects in CA literature) that have common properties (referred to as attributes in CA literature). These groupings are called concepts and capture similarities among a set of elements based on their common properties. In the specific case of software reengineering, the system are composed of a big amount of different entities (classes, methods, modules, subsystems) and there are different kinds of relationships among them. It also represents dependencies among the classes or entities.

X-Ray views —a technique based on Concept Analysis— which reveal the internal relationships between groups of methods and attributes of a class. X-Ray views are composed out of elementary collaborations between attributes and methods and help the engineer to build a mental model of how a class works internally.

2. Existing Idea

Within object oriented software, the minimal unit of development and testing is a class. Usually, a class is composed of instance variables used to represent the state, and methods used to represent the behavior of the classes. Then, understanding how a class works means identifying several aspects:

How the methods are interacting together (coupling between methods)

How the instance variables are working (or not) together in the methods (coupling between instance variables)

Which methods are using (or not) the state of the class

if there are methods that form a cluster and define together a precise behaviour of the class

Which methods are considered as interfaces

Which methods are used as entry points (methods that are considered as interfaces and communicate with other methods defined in the class)

Which methods and instance variables represent the core of the class

Which methods are using all the state of the class

In paper [1], the authors have given an idea of concept analysis. Mathematically, concepts are maximal collections of elements sharing common properties. To use the CA technique, one only needs to specify the properties of interest on each element, and does not need to think about all possible combination of these properties, since these groupings are made automatically by the CA algorithm. The possibility of capturing similarities of elements in groups (concepts) -based on the specification of simple properties allow to identify common features of the elements. When we are able to characterize the entities in terms of properties, and we can detect if these characteristics are repeated in the system, then we can reduce the amount of information to analyze and we can have an abstraction of the different parts of a system. These abstractions help us to start to see how the parts are working, how they are defined and how they are connected to other parts of the system .The elements are the instance variables and the methods defined in the class, and the properties are how they are related between themselves.

If we have the set of instance variables {A, B}, and the set of methods {P, Q, X, Y} defined in a class, the properties we use are:
B is used by P means that the method P is accessing directly or through an accessor / mutator to the instance variable B.
Q is called in P means that the method Q is called in the method P via a self-call. It also shows indirect dependencies between elements if exists. They have also shown different types of relations and dependencies through some notations.

{E1, ..,En} R {M1, ..,Mp} means that the entities {E1, ..,En} depend exclusively on {M1, ..,Mp}. This means that {M1, ..,Mp} are the only entities that are related through the property R to {E1, ..,En}.

{E1, ..,En} R {M1, ..,Mp} means that the entities {E1, ..,En} do not depend exclusively on {M1, ..,Mp}.

{E1, ..,En} $\overline{R}$*{M1, ..,Mp} means that the entities{E1, ..,En} depend exclusively and transitively on {M1, ..,Mp}. This means that {M1, ..,Mp} are the only ones that are related to {E1, ..,En} through the property R and R1, where R1 is an intermediate property, because there is a set {N1, ..,Nk} such that: {E1,..,En} R {N1, ..,Nk} R1 {M1, ..,Mp}.

{E1, ..,En} R□ {M1, ..,Mp} means that the entities {E1, ..,En} do not depend exclusively but transitively on {M1, ..,Mp}. This means that {M1, ..,Mp} are not the only ones that are related to {E1, ..,En} through the property R and R1, where R1 is an intermediate property, because there is a set {N1, ..,Nk} such that: {E1, ..,En} R {N1, ..,Nk} R1 {M1, ..,Mp}.

A special case: {E1, ..,En} ¬R {M1, ..,Mp} means that the entity {E1, ..,En} has any dependencies on {M1, ..,Mp}. This is only applicable on exclusive dependencies.

In paper [2], the authors have discussed different types of X-ray views which will be helpful for our future work. In paper [3] there is a concept on modularization using the conceptual analysis on object oriented environment.

3. Our Application

Our idea is now to use the above said concepts in the environment of java and to way out the modularization in java programs. Modularization also helps in software re-engineering.
For the present purpose, let us have an example of java coding. We have applied the proposed idea in different properties of Java programming each of which are illustrated below.

3.1 Polymorphism

Polymorphism deals with of different forms of a method where parameters are different according to the forms of the methods. Polymorphism can also occur in constructors.

```
class Overload {
int  a;
void test(int x) {
a=x;
System.out.println("a: " + a);
}
void test(int x , int y) {
a= x;
int b= y;
System.out.println("a and b: " + a + "," + b);
}
}

class MethodOverloading {
public static void main( String args[ ] ) {
Overload overload = new Overload( );
overload. test(10);
overload. test(10, 20);

}
}
```

instance variable a is not mutually related to test(int x) or test(int x, int y) as both are accessing the variable.

{ a } $\overline{R}$ { test(int x), test(int x, int y) } --------------- 1

{ b } R { test(int x, int y) }-------------------------------2

So, where 2 relations are found and we can say that these two relations will create two concepts.

3.2 Overriding

Overriding is runtime polymorphism. The methods are same in syntax. It is decided in the run time which method is to be invoked.

Method overriding.

```
class A {
int  i, j;
A( int a, int b ) {
i = a;

 j = b;

}


// display i and j
void show( )

{
System.out.println("i and j: " + i + " " + j );
}
}

class B {
int k;
B( int c) {
k = c;
}

void show( ){ // display k – this overrides show( ) in
A

System.out.println("k: " + k);
    }
  }

class Override {
public static void main(String args[ ] ) {
B subOb = new B(1 );
subOb.show( ); // this calls show( ) in B
}
}
```

A( ) accessing the instance variables i , j

A .Show( ) [show( ) of class A] accessing the instance variables directly for both A( ) and A. show( ) the relation comes like :

{i, j} $\overline{R}$ { A( ), A. show( ) } --------------------- 1

Similarly B( ) and B. show( ) exclusively related to variable k. so we can say that the relations are like this:

{B( ), B. Show( )} $\overline{R}$ {k} ------------------------ 2

So , two relations are creating two concepts. Now as the method show( ) is overridden, we shall consider the A. show( ) and B. show ( ) as a single entity say show( ). As we are considering here only the property of overriding we shall ignore the other methods and we can reduce the relations or dependencies like :

show( ) $\overline{R}$ { i, j, k } and therefore creating a module.

3.3 Inheritance

Inheritance is a property in Java where the members of a class inherit properties or attributes from its base classes. Inheritance can be of different forms multiple, hierarchical, multistage and hybrid.

```
 class A {
       int x;
       int y;
       void showxy ( )

 {

System.out.println("  x and y :  " + x + " " + y );
   }
   }

class B extends A

{

int z;
  void showz( ){
    System.out.println( "z : " + z );
   }

}


class Inheritance{
public static void main(String args[ ] )

{
   A a = new A( );

B b= new B( );
    a . x = 5;        // x of superclass A


     a.y = 5;          // y of superclass A
showxy ( );  // showxy ( ) of class A i.e. the superclass

       x= 10;      // x of subclass B as extended from A

       y= 10;     // y of subclass B as extended from A

       k= 10;     // k of own subclass B

  b.showxy( ); // showxy of superclass A extended by

                           subclass B

       b. showz( );   // own method showz( ) of class B
           }
       }
```

x, y is mutually exclusively related to {showxy ( )} in case of class A and in case of class B too because showxy ( ) is inherited by class B from class A. So we can say that :

{x, y } $\overline{R}$ { showxy( ) }. Here one concept is created.

---------------------------- 1

Next, showz ( ) is accessing mutually exclusively to z. So, the relation goes like this :

{z } $\overline{R}$ {showz( )}. Another concept is created.--------- 2

As all relations are mutually exclusive we can take aggregation and can be written as:

{x, y, z }$\overline{R}${ showxy( ), showz( ) }----------------------- 3

3.4 Exception Handling

Exception handling is the property of java by which it can invoke some work when some normal task is prevented to execute by some faulty codes.

```
Class MyException {

       public static void main ( String args [ ] ) {
```

```
int d, a;

try {

d= 0;



a= 42 / d;

System. Out. Println ( “ This will not be printed. “);

} catch (ArithmeticException e ) {  // catching of
divided by zero errors

System.out.println( “ This is Division by zero creating
an exception !!!” );

}  System.out.println( “ This is after the catch
done….” );

            }

}
```

In the class MyException a and d are instance variables. We can say that, try-catch block is directly accessing the variables because whenever the try block is executed then only the ArithmeticException e arises i.e. the instance variables are mutually exclusive with exception e. Thus they are creating concepts and therefore a module. By notation we write that :

{ a, d} $\overline{\text{R}}$ {try-catch ( ) }

3.5 Abstraction

Abstraction is the property of Java to hide details from the users so that the user can deal only with the functionalities of the codes.

```
 class Poly
{
// implementations and private members hidden

Poly (int , int );
double eval ( double );
void add ( Poly );
void mult ( Poly );
public String toString ( );
}

public class Binomial
{
public static void main ( String[ ] args )

{

int N = Integer.parseInt ( args [ 0 ] ) ;

double p = Double. parseDouble(args [1]) ;

Poly y = new Poly ( 1, 0);
Poly t = new Poly ( 1, 0);
t.add ( new Poly ( 1, 1) );
for ( int i = 0; i < N; i++ )
{
y. mult ( t ) ;
Out.println(y + "");
}
Out.println("value: " + y.eval ( p ) );
}
}
```

Method add ( ) is using poly ( ) constructor. So, by notation we can write that :

{ add ( ) } $\overline{\text{R}}$ * { poly ( ) } ----------------------- (1)

Moreover, the method add( ) and mult ( ) using the total class methods directly or indirectly as we can see from the class definition .Then also we can write that,

{add ( ), mult ( ) } $\overline{\text{R}}$* {class Poly } ------------- (2)

It means that those methods are using the class methods otherwise.

4. Conclusion

We have tried to implement the basic properties of object oriented paradigm through concept analysis notation to create concepts as well as the modules. These modules will help us for reengineering because reengineering deals with change in the modules of codes to make the obsoluted or about to obsoluting software rework.

Acknowledgment

All complements to the almighty who gives us the ability and knowledge to do this hard work. We express our greatest gratitude, appreciation and profound respect to our supervisor G Geetha, Dept. of Computer Applications, Lovely Professional University, Jalandhar, Punjab (India) for his continuous, proper & perfect guidance and invaluable suggestion.

References

[1] Gabriela Ar´evalo, Stephan Ducasse, Oscar Nierstrasz. “X-ray view on a Class using Conceptual Analysis” published in the Conference at University of Antwerp, p: 76-80 in 2003.

[2] Gabriela Ar´evalo, Stephan Ducasse, Oscar Nierstrasz. “Understanding Classes using X-ray views” cited in the Proceedings of 2nd. MASPEGHI (ASE), p: 9-18 in 2003.

[3] H. H. Kim, Doo-Hwan Bae. “Object-oriented concept analysis for software modularization” cited in the Proceedings of IET Software, p: 134~148 in 2008.

[4] S. Demeyer, S. Ducasse, and O. Nierstrasz. Object-Oriented Reengineering Patterns. Morgan Kaufmann, 2002.

[5] B. Ganter and R. Wille. Formal Concept Analysis: Mathematical Foundations. Springer Verlag, 1999.

[6] M. Fowler. Refactoring: Improving the Design of Existing Programs. Addison-Wesley, 1999.

Authors’ Profile

Gulshan Kumar pursuing his M. Tech degree in Computer Science and Engineering from Lovely Professional University, Jalandhar, India. His research interest includes Cryptography and Mobile Adhoc Networks and Software Engineering.

Mritunjay Kumar Rai received his Ph.D. Degree from ABV-Indian Institute of Information Technology and Management, Gwalior, India. Currently he is working as an Assistant Professor in Lovely Professional University. His research interest area is Mobile Adhoc Networks and Wireless Sensor Networks.

A fast multi-class SVM learning method for huge databases

Djeffal Abdelhamid[1], Babahenini Mohamed Chaouki[2] and Taleb-Ahmed Abdelmalik[3]

**[1,2] Computer science department, LESIA Laboratory,
Biskra University, Algeria**

[3] LAMIH Laboratory FRE CNRS 3304 UVHC, Valenciennes university, France

Abstract

In this paper, we propose a new learning method for multi-class support vector machines based on single class SVM learning method. Unlike the methods 1vs1 and 1vsR, used in the literature and mainly based on binary SVM method, our method learns a classifier for each class from only its samples and then uses these classifiers to obtain a multiclass decision model. To enhance the accuracy of our method, we build from the obtained hyperplanes new hyperplanes, similar to those of the 1vsR method, for use in classification. Our method represents a considerable improvement in the speed of training and classification as well the decision model size while maintaining the same accuracy as other methods.

Keywords: *Support vector machine, Multiclass SVM, One-class SVM, 1vs1, 1vsR.*

1. Introduction and related work

The support vector machine (SVM) method is, in its origin, binary. It is based on the principle of separation between the samples of tow classes, one positive and one negative.

In this form, the SVM method is very successful in several areas of application given the precision it offers. In practice, we find more applications with multi-class problems, hence the need to extend the binary model to meet multi-class problems. Existing methods currently try mainly to optimize two phases: a training phase and a classification phase. The first phase constructs the hyperplane, and the second uses it.

The evaluation of the methods is based on the evaluation of the performances of the two phases. Among the well known methods, there are methods for direct solution without using the binary SVM model as Weston \& Watkins model [1], but which suffers, always, from some slowness and weak accuracy. The widely used methods are based essentially on the extension of the binary model, namely, the one-against-rest (1vsR) and the one-against-one (1vs1) methods. The 1vsR method learns for each class a hyperplane that separates it from all other classes, considering this class as positive class and all other classes as negative class, and then assigns a new sample, in the classification phase, to the class for which it maximizes the depth. The 1vs1 method learns for each pair of classes a separating hyperplane, and uses the voting lists or decision graphs (DAG) to assign a new sample to a class [2,3].

In [4,5,6,7] comparative studies are conducted to assess the performances of these methods. According to the authors, the 1vs1 method is faster while the 1vsR method is more accurate. \\

In this paper, instead of the binary SVM, we propose to use the one-class SVM (OC-SVM) that provides a hyperplane for each class that separates it from the rest of the space. For each class, we learn a hyperplane from only its samples. Then in the classification phase, we build for each class a new two-class hyperplane which separates it from all other classes. This two-class hyperplane is calculated from the previous one-class hyperplane and the closest sample of the other classes. The new two-class hyperplane is a shift of the one-class hyperplane, it is situated between the farthest misclassified sample, of the target class, from the hyperplane, and nearest sample, belonging to other classes, to the hyperpalne. Our technique speeds up the training time and classification time, and reduces the decision model size compared to classic methods, while keeping very close accuracy. Our results were validated on toys and then on databases of UCI site [8]. The rest of the paper is organized as follows: section 2 introduces the binary SVM and Section 3 introduces the multi-class methods based on the binary model, namely, 1vsR and 1vs1. In section 4, we present the OC-SVM model and then we present our method in Section 5. The results and their discussion are presented in Section 6, and a conclusion in Section 7.

2. Binary SVM

The binary SVM solves the problem of separation of two classes, represented by n samples of m attributes each

[9,10]. Consider the problem of separating two classes represented by n samples:

$$\{(x_1, y_1), .., (x_n, y_n)\}, x_i \in \Re^m, y_i \in \{-1, +1\}\}$$

Where x_i are learning samples and y_i their respective classes. The objective of the SVM method is to find a linear function *f* (equation 1), called *hyperplane* that can separate the two classes:

$$\begin{cases} f(x) = (x \bullet w) + b; & \\ f(x) > 0 & \Rightarrow x \in class + 1 \\ f(x) < 0 & \Rightarrow x \in class - 1 \end{cases} \tag{1}$$

Where *x* is a sample to classify, *w* is a vector and b is a bias.

We must therefore find the widest margin between two classes, which means minimizing w^2. In cases where training data are not linearly separable, we allow errors ξ_i (called slack variables) of samples from boundaries of the separation margin with a penalization parameter C and the problem becomes a convex quadratic programming problem:

$$\begin{cases} Minimize & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \\ under\ constraints & \\ & y_i(w^T x_i + b) \geq 1 - \xi_i; i = 1..n \\ & \xi_i \geq 0 \end{cases} \tag{2}$$

The problem of equation 2 can be solved by introducing Lagrange multipliers α_i in the following dual problem:

$$\begin{cases} Minimize & \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i x_j) - \sum_{i=1}^{n}\alpha_i \\ under\ constraints & \\ & \sum_{i=1}^{n}\alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C \end{cases} \tag{3}$$

Hence, we can have the following decision function (hyperplane):

$$H(x) = \sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b \tag{4}$$

The function *K* is called *Kernel*, it is a symmetric function that satisfies Mercer conditions [4]. It can represent a transformation of the original input space in which data could be non-linearly separable to a new larger space where a linear separator exists. Solving the problem of equation 3 requires an optimization, especially when the number of samples is high. Among the optimization methods most commonly used, there is the SMO (Sequential Minimal Optimization) where the problem is broken into several sub-problems, each optimizes two α_i [11].

3. Multi-class SVM

Several techniques have been developed to extend binary SVM method to problems with multiple classes. Each of these techniques makes a generalization of the abilities of the binary method to a multi-class field [2,3]. Among the best known methods, we can cite 1vsR and 1vs1.

3.1 One-against-rest (1vsR)

For each class *k* we determine a hyperplane $H_k(w_k, b_k)$ separating it from all other classes, considering this class as positive class *(+1)* and other classes as negative class *(-1)*, which results, for a problem of *K* classes, to *K* binary SVMs. A hyperplane H_k is defined by the following decision function:

$$f_k(x) = \langle w_k \bullet x \rangle + b_k \tag{5}$$

This function allows to discriminate between the samples of the class *k* and the set of all other classes.

Classification: A new sample *x* is assigned to the class k^* that maximizes the depth of this sample. This class is determined by the decision rule of the equation 6.

$$k^* = Arg_{(1 \leq k \leq K)} Max f_k(x) \tag{6}$$

Figure 1 shows an example of separation of three classes.

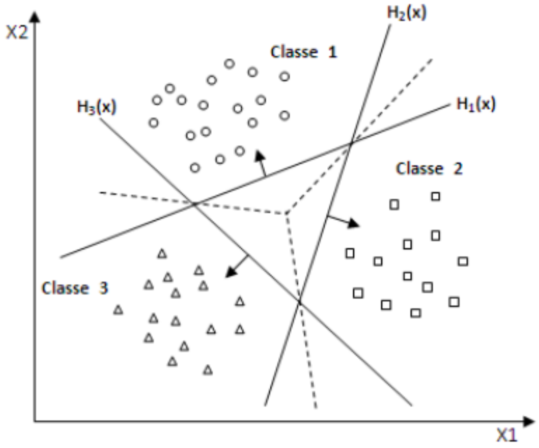


Fig. 1 One-against-rest approach

Interpreted geometrically, a new sample *x* is assigned to the class that corresponds to the farthest hyperplane. Thus, the space is divided into *K* convex regions, each corresponding to a class.

3.2 One-against-one (1vs1)

Training: Instead of learning *K* decision functions, the 1vs1 method discriminates each class from every other class. Thus *K(K-1)/2* decision functions are learned. The one-against-one approach is, thus, to build a classifier for each pair of classes *(k,s)*.

Classification: For each pair of classes *(k, s)*, the 1vs1 method defines a binary decision function H_{ks}: $x \in \backslash\Re^{m} \rightarrow \{-1,+1\}$. The assignment of a new sample can be done by two methods; voting list or decision graph.

a.Voting list: We test a new sample by calculating its decision function for each hyperplane. For each test, we vote for the class to which the sample belongs. We define, for this, the binary decision function $H_{ks}(x)$ of equation 7.

$$H_{ks}(x) = sign(f_{ks}(x)) = \{+1 \quad if \quad f_{ks}(x) > 0; \quad 0 \quad else\} \qquad (7)$$

Based on the *K (K-1)/2* binary decision functions, we define other *K* decision functions (equation 8):

$$H_k(x) = \sum_{s=1}^{m} H_{ks}(x) \qquad (8)$$

The classification rule of a new sample *x* is given, then, by the equation 9:

$$k^* = Arg_{(1 \le k \le K)} Max H_k(x) \qquad (9)$$

Figure 2 is an example of classification of three classes.

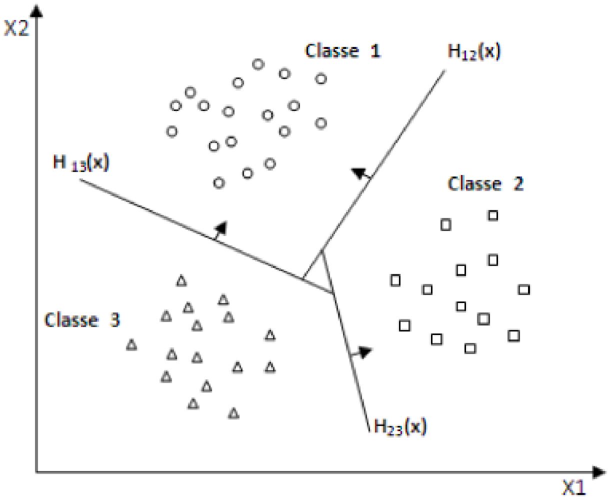


Fig. 2 One-against-one approach

b. Decision graphs: In this method, we define a measure E_{ks} of the generalization ability of the different obtained hyperplanes i.e for each pair of classes. This measure (equation 10) represents the ratio between the number of support vectors of the hyperplane and the number of samples of both classes.

$$E_{ks} = \frac{N_{vs}}{N_{exemples}} \qquad (10)$$

Before deciding about new samples, we construct a decision graph. We start with a list *L* containing all the classes, then take the two classes *k* and *s* which E_{ks} is maximum, and we create a graph node labeled *(k,s)*. We then create, in the same way, a left son of this node from the list *L-{k}* and a right son from the list *L-{s}*, and continue until the list *L* contains only one class. This gives the decision graph of figure 3, whose leaves are the classes and the internal nodes are the hyperplanes:

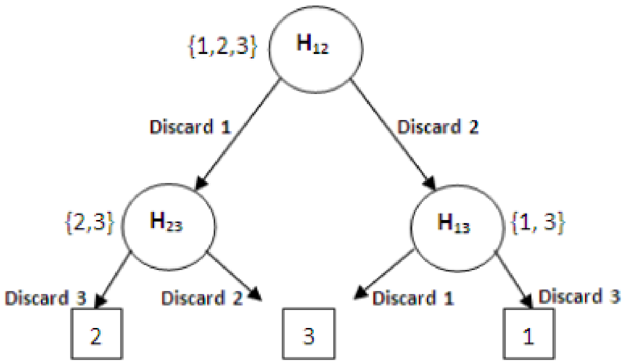


Fig. 3 Directed acyclic decision graph for three classes

A new sample *x* is exposed first to the hyperplane of the root, if the decision is positive then we continue with the left son, otherwise with the right son, until a leaf is reached. The reached leaf represents the class of the sample *x*.

4. One-class SVM

In the one-class SVM classification, it is assumed that only samples of one class, the target class, are available. This means that only the samples of the target class can be used and no information on other classes is present. The objective of the OC-SVM is to find a boundary between the samples of the target class and the rest of space. The task is to define a boundary around the target class, so that it accepts as many target samples as possible [12]. The one-class SVM classifier considers the origin as the only instance of negative class, and then tries to find a separator between the origin and samples of the target class while maximizing the margin between the two (cf. figure 4).

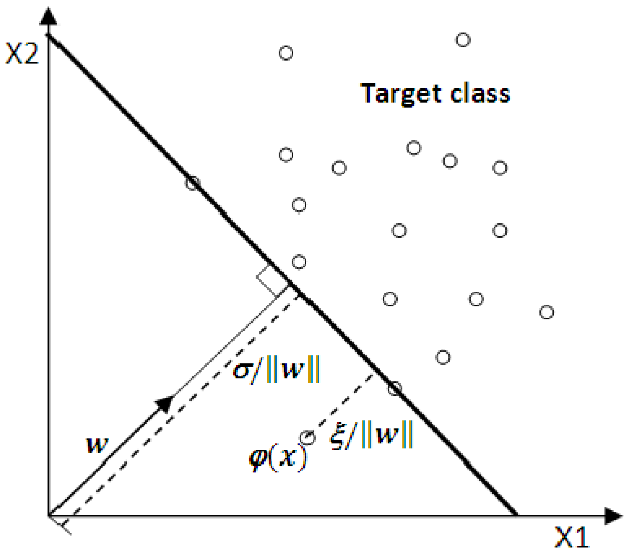


Fig. 4 One-class SVM with maximum margin

The problem is to find the hyperplane that separates the target samples of the origin while maximizes the distance between them. The problem is modeled by the primal quadratic programming problem of equation 11.

$$\begin{cases} min_{w,\xi,\rho}\frac{1}{2}\left\|w\right\|^2 + \frac{1}{vN}\sum_{i=1}^{l}\xi_i - \rho \\ \langle w, \phi(x_i)\rangle \geq \rho - \xi_i \\ \xi_i \geq 0 \qquad i = 1, 2..l \end{cases} \tag{11}$$

Where *N* is the number of samples of the target class, *(w,ρ)* parameters to locate the hyperplane, ζ_i the allowed errors on samples classification, penalized by the parameter *v* and φ is a transformation of space similar to the binary case. Once *(w,ρ)* determined, any new sample can be classified by the decision function of equation 12:

$$f(x) =< w, \phi(x_i) > -\rho \tag{12}$$

x belongs to the target class if *f(x)* is positive. In fact, solving the problem of the equation 11 is achieved by the introduction of Lagrange multipliers in the dual problem of equation 13:

$$\begin{cases} Minimize_\alpha & \frac{1}{2}\sum_{i,j}\alpha_j K(x_i, x_j) \\ under\ constraints & \\ & \sum_{i=1}^{n}\alpha_i = 1 \\ & 0 \leq \alpha_i \leq \frac{1}{vl} \end{cases} \tag{13}$$

Where *K* is a kernel that represents the space transformation φ. Once the α_i are determined using an optimization such as SMO [11], the decision function for any sample *x* is given by equation 14:

$$f(x) = \sum_{i=1}^{l}\alpha_i K(x_i, x) - \rho \tag{14}$$

Where ρ can be determined from a training sample x_i having $\alpha_i \neq 0$ by the equation 15:

$$\rho = \sum_j \alpha_j K(x_j, x_i) \tag{15}$$

5. Multi-class SVM method based on OC-SVM (OCBM-SVM)

5.1 Training

The method we propose in this paper, extends the OC-SVM to multi-class classification. We propose to learn for each class its own hyperplane separating it from the rest of the space. We learn so for *K* classes, *K* hyperplane, but unlike the 1vsR method, we use to find each hyperplane H_k only the samples of the class *k* which speeds up considerably the training time. Table 1 shows the cost of different methods in terms of the number of used hyperplanes, the number of samples used by each hyperplane, and the estimated time of training depending on the number of samples of a class N_c. We assume, for the sake of simplicity, that the classes have the same number of samples. The estimated time is based on the fact that each hyperplane H_k using N_k samples requires a time of βN_k^2 (β represents the conditions of implementation such as CPU speed, memory size,...etc.).

Table 1: Comparative table of the training times of the different methods

Method	# Hyperplanes	# samples/Hyperplane	Estimated time
1vsR	K	KN_c	$K^3\beta N_c^2$
1vs1	$K(K-1)/2$	$2N_c$	$2\beta K^2 N_c^2$
OCBM-SVM	K	N_c	$K\beta N_c^2$

- The 1vsR method uses to determine each of the *K* hyperplanes all KN_c training samples, which leads to a training time of $K^3\beta N^2$, where *K* represents the number of classes and β a constant related to the conditions of implementation, based on the fact that the SMO algorithm [11] used here is of complexity $O(N^2)$.
- The 1vs1 method calculates a hyperplane for each pair of classes, i.e *K(K-1)/2* hyperplanes, and to determine a hyperplane separating two classes, we use the samples of these two classes $2N_c$, resulting in a training time of $2\beta K^2 N_c^2$.

- Our method requires the calculation of K hyperplanes, each separates a class from the rest of space. To determine each hyperplane, we use only the samples of one class, resulting therefore in a total training time of about $2KN_c^2$.

It is clear that:

$$K\beta N_i^2 < 2\beta K^2 N_i^2 < K^3 \beta N_i^2 \qquad (16)$$

This means that the proposed method optimizes the training time compared to methods 1VsR and 1vs1.

5.1 Classification

OC-SVM method allows to find a hyperplane separating one class from the rest of space, this hyperplane allows to decide on membership of a new sample to this class. If the sample is above the hyperplane then it belongs to the class, but if it is below, we have no information on to what other class the sample belongs. To correct this situation, we propose to modify the obtained hyperplane to enhance the decision information in the case where the sample does not belong to the class. We propose to find for each hyperplane of a class, the closest sample among the samples of all other classes, then shift the hyperplane by a distance equal to half the distance between this sample and the misclassified sample, of the target class, that minimizes the decision function (the farthest one from the hyperplane) (cf. Figure 5).

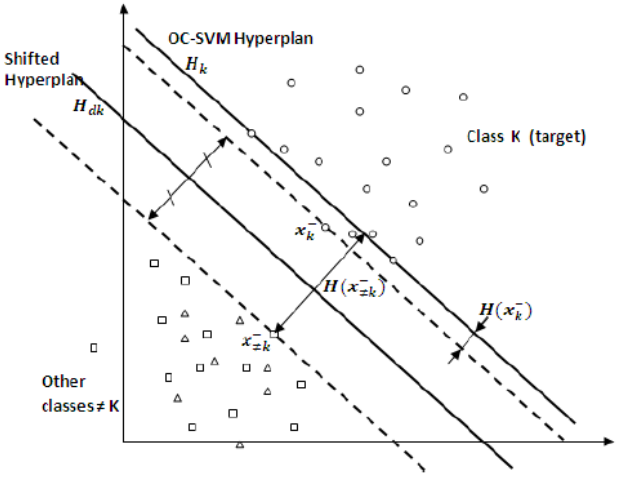


Fig. 5 Classification in OCBM-SVM method

More formally, let $H_k(x)$ be the decision function using the hyperplane of the kth class. And let x_k^- be the misclassified sample of the class k farthest from the hyperplane H, and $x_{\neq k}^-$ the closest sample to the hyperplane H, belonging to a class different to k. The distance of x_k^- from the hyperplane is given by $H_k(x_k^-)$ and the distance between the hyperplane and $x_{\neq k}^-$ is given by $H_{\neq k}(x_k^-)$. The proposed shift is $[H_k(x_k^-) + H_k(x_{\neq k}^-)]/2$, and the new decision function for a sample x can be calculated by the shifted hyperplane Hdk of the equation 17:

$$H_{dk}(x) = H_k(x) - \frac{(H_k(x_k^-) + H_k(x_{\neq k}^-))}{2} \qquad (17)$$

After calculating all the K shifted hyperplanes, the decision about the class k^* of a new sample x can be given by the discrete decision rule of equation 18:

$$k^* = Arg_{(1 \leq k \leq K)} max H_{dk}(x) \qquad (18)$$

Table 2 shows a comparison between the time of classification of a new sample by different methods. This time is calculated based on the number of support vectors of each hyperplane, which is equal, to the maximum, the total number of samples used to find the hyperplane.

Table 2: Comparative table of classification time of the different methods

Method	# Used hyperplanes	Estimated time
1vsR	K	$K^2\beta N_c$
1vs1	$K(K-1)/2$	$K(K-1)\beta N_c$
DAG	$(K-1)$	$2(K-1)\beta N_c$
OCBM-SVM	K	$K\beta N_c$

- The 1vsR method uses K hyperplanes, and to test a sample, it is evaluated for all hyperplanes. Knowing that each hyperplane contains a number of support vectors equal to the maximum total number of samples, the estimated time to find the class of a sample is $K^2\beta N_c$, where β is a constant that represents the conditions of implementation.
- For the method 1vs1 using the voting list, the classification estimated time is $K(K\text{-}1)\beta N_c$,
- For the method 1vs1 using decision graphs, the classification estimated time is $2(K\text{-}1)\beta N_c$.
- The estimated time for our method is $K\beta N_c$, because it tests the sample for K hyperplanes containing, at maximum, KN_c support vectors.

If we eliminate from the column of estimated time in Table 2 the factor βN_c, we note that the method OCBM-SVM optimizes the classification time. But this depends always on the number of support vectors obtained by the training method which can be very important compared to the number of classes K. Higher the number of training samples and the number of classes, the greater the advantage of the OCBM-SVM method, since it uses the minimum of samples, making it suitable for large databases.

5.3 Model size

The method we propose in this paper, extends the OC-SVM to multi-class classification. We propose to learn for each class its own hyperplane separating it from the rest of the space. We learn so for K classes, K hyperplane, but unlike the 1vsR method, we use to find each hyperplane H_k only the samples of the class k which speeds up considerably the training time. Table 1 shows the cost of different methods in terms of the number of used hyperplanes, the number of samples used by each hyperplane, and the estimated time of training depending on the number of samples of a class N_c. We assume, for the sake of simplicity, that the classes have the same number of samples. The estimated time is based on the fact that each hyperplane H_k using N_k samples requires a time of βH_k^2 (β represents the conditions of implementation such as CPU speed, memory size,...etc.).

6. Experiments

6.1 Used data

Our method was first tested on examples of Toy type (2 dimensions) that we have chosen of different complexities. Then we have tested it on benchmarks of multi-class classification databases from the site "Machine Learning Repository UCI" [8]. The used databases are shown in the following Table 3:

Table 3: Databases used for testing

Base	Domain	N_{att}	N_{class}	N_{train}	N_{test}
PageBlocks	Web	10	5	389	5090
Segment	Image processing	19	7	500	1810
Vehicule	Industry	18	4	471	377
Abalone	Industry	8	27	3133	1044
Letter	Character recognition	16	26	2000	10781
OptDigits	Industry	64	10	3824	1797

N_{att} is the number of attributes, N_{class} is the number of classes in the database, N_{train} is the number of samples used for training and N_{test} is the number of samples used for testing.

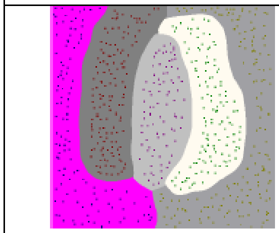
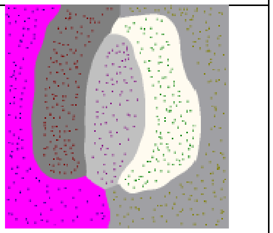
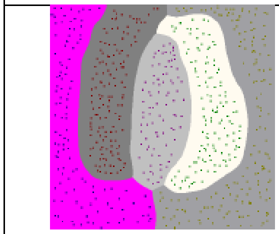
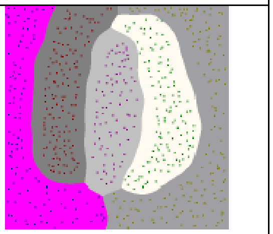
6.2 Materials and evaluation criteria

The proposed method was tested on a Dual-core de 1.6 GHZ machine with 1GB of memory. The used kernel is RBF, and the optimization method is the SMO algorithm [11]. The evaluation criteria of the performances of our method are the training time in seconds $T_r(s)$, the classification time in seconds $T_c(s)$, the recognition rate R and the size of the obtained model.

6.3 Results

The first tests have been performed on toys of different complexities and have shown the advantages of OCBM-SVM method compared to other metods. In the example of table 4, we took 642 samples belonging to five classes. The results show that our OCBM-SVM method reduced training time to 0.109 second without losing accuracy, while the 1vs1 and 1vsR methods have given respectively 0.609 and 28.25 seconds. Even, the size of the resulting model was reduced to 18.290 KB. The classification time was 22.625 which is close to that of the method of decision graphs (DAG) with the remark that the number of samples and the number of classes are small.

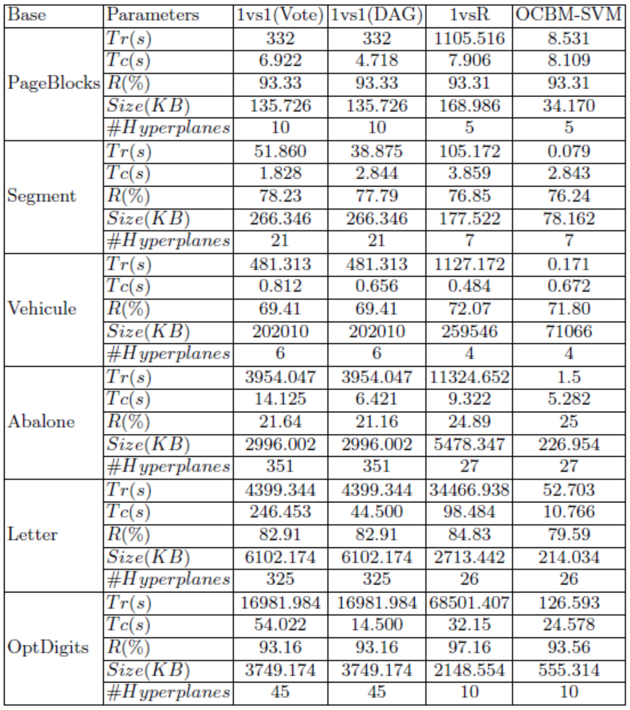
Table 4: Results on a toy

1vs1				DAG			
							
Tr(s)	Tc (s)	R(%)	Size(KB)	Tr(s)	Tc (s)	R(%)	Size(KB)
8.609	41.922	100	24.71	8.609	20.610	100	24.71
1vsR				OCBM-SVM			
							
Tr(s)	Tc(s)	R(%)	Size(KB)	Tr(s)	Tc(s)	R(%)	Size(KB)
28.250	44.890	100	23.714	0.109	22.625	100	14.282

The tests performed on the databases of the UCI site also confirm the theoretical results. Table 5 summarizes the results obtained on testing databases presented in the table 3.

Indeed, in all tested databases, OCBM-SVM method greatly improves the training time and model size, especially in large databases. For the *Abalone* database, our OCBM-SVM method reduced the training time from 3954.047 seconds to 1.5 seconds and the size of the model from 2996,002 KB to 226,954 KB. In the case of *OptDigits* database, training time was reduced from 16981.984 seconds for 1vs1 method and 68501.407 seconds (over 19 hours) for 1vsR method, to only 126,593 seconds while maintaining accuracy better than the method 1vs1. For the same database, the size of the model was also reduced from 3749.174 KB for the 1vs1 method and 2148,554 KB for 1vsR to 555,314 KB.

Table 5: Results on different databases



Base	Parameters	1vs1(Vote)	1vs1(DAG)	1vsR	OCBM-SVM
PageBlocks	$Tr(s)$	332	332	1105.516	8.531
	$Tc(s)$	6.922	4.718	7.906	8.109
	$R(\%)$	93.33	93.33	93.31	93.31
	$Size(KB)$	135.726	135.726	168.986	34.170
	$\#Hyperplanes$	10	10	5	5
Segment	$Tr(s)$	51.860	38.875	105.172	0.079
	$Tc(s)$	1.828	2.844	3.859	2.843
	$R(\%)$	78.23	77.79	76.85	76.24
	$Size(KB)$	266.346	266.346	177.522	78.162
	$\#Hyperplanes$	21	21	7	7
Vehicule	$Tr(s)$	481.313	481.313	1127.172	0.171
	$Tc(s)$	0.812	0.656	0.484	0.672
	$R(\%)$	69.41	69.41	72.07	71.80
	$Size(KB)$	202010	202010	259546	71066
	$\#Hyperplanes$	6	6	4	4
Abalone	$Tr(s)$	3954.047	3954.047	11324.652	1.5
	$Tc(s)$	14.125	6.421	9.322	5.282
	$R(\%)$	21.64	21.16	24.89	25
	$Size(KB)$	2996.002	2996.002	5478.347	226.954
	$\#Hyperplanes$	351	351	27	27
Letter	$Tr(s)$	4399.344	4399.344	34466.938	52.703
	$Tc(s)$	246.453	44.500	98.484	10.766
	$R(\%)$	82.91	82.91	84.83	79.59
	$Size(KB)$	6102.174	6102.174	2713.442	214.034
	$\#Hyperplanes$	325	325	26	26
OptDigits	$Tr(s)$	16981.984	16981.984	68501.407	126.593
	$Tc(s)$	54.022	14.500	32.15	24.578
	$R(\%)$	93.16	93.16	97.16	93.56
	$Size(KB)$	3749.174	3749.174	2148.554	555.314
	$\#Hyperplanes$	45	45	10	10

The proposed method keeps a classification time, in the case of large databases, close to that of the method DAG representing the fastest method in terms of classification time. Indeed, we note, in databases with a number of samples less than 1000 (case of *PageBlocks*, *Segment* and *Vehicule*) that the classification time obtained by some methods is better than ours. This is due to that in the databases of small size, the preparation work of classification structures (lists in DAG and shifts in OCBM-SVM) is important compared to the essential task of calculating the decision functions. In large databases, with number of samples greater than 1000 and number of classes greater than 10 (case of Abalone, *Letter*, *OptDigits*), the improvement of classification time is remarkable. Indeed, for the Abalone database, our method yielded a classification time of 5.282 seconds against 6.421 seconds for DAG method the best of the others.

Also for the Letter database, the classification time was 10,766 seconds against 44.5 seconds for DAG. In the *OptDigits* database, the obtained time by the DAG method was better than the OCBM-SVM method, this can be explained by the number of support vectors obtained by each training method which has an important influence on the classification time.

With all its advantages, our method preserves a recognition rate in the range of rates obtained by other methods and sometimes better (case of *Abalone* database).

7. Conclusion

In this paper, we presented a new method for multiclass learning with support vector machine method. Unlike classic methods such as 1vs1 and 1vsR extending the principle of binary SVM, our method (denoted OCBM-SV) extends the principle of one-class SVM method. And to achieve the generalization ability of binary SVM, we changed the hyperplanes obtained by the one-class method to take into account information of other classes. The obtained results show great improvement in training time and decision model size. Our method also allows improving the classification time (decision making) of new samples by reducing of the number of support vectors of the decision model. The obtained accuracy is very close to that of other methods and sometimes better.

8. References

[1] U. Dogan, T. Glasmachers, C. Igel, "Fast Training of Multi-class Support Vector Machines", Technical Report no. 03/2011, Faculty of science, university of Copenhagen, 2011.

[2] S. Abe, "Analysis of multiclass support vector machines", In Proceedings of International Conference on Computational Intelligence for Modelling Control and Automation (CIMCA'2003), pp. 385-396, 2003.

[3] Y. Guermeur, "Multi-class Support vector machine, Theory and Applications", HDR thesis, IAEM Lorraine, 2007.

[4] Y. Liu, R. Wang, Y. Zeng, H. He, "An Improvement of One-against-all Method for Multiclass Support Vector Machine", 4th International Conference: Sciences of Electronic, Technologies of Information and telecommunications, March 25-29, 2007, TUNISIA.

[5] N. Seo, "A Comparison of Multi-class Support Vector Machine Methods for Face Recognition", Research report, The University of Maryland, Dec 2007.

[6] G. Anthony, H. Gregg, M. Tshilidzi, "Image Classification Using SVMs: One-against-One Vs One-against-All", Proccedings of the 28th Asian Conference on Remote Sensing, 2007.

[7] M. G. Foody,A. Mathur, "A Relative Evaluation of Multiclass Image Classification by Support Vector Machines", IEEE Transactions on Geoscience and Remote Sensing, V. 42 pp. 13351343, 2004.

[8] A. Frank, A. Asuncion, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml], Irvine, CA: University of California, School of Information and Computer Science, 2010.

Fast FPGA Implementation of EBCOT block in JPEG2000 Standard

Anass Mansouri, Ali Ahaitouf, and Farid Abdi

UFR SSC, LSSC, Electrical Engineering Department Faculty of sciences & technology BP: 2202 FES MOROCCO
Phone: +212 35 61 13 26, Fax: +212 35 60 82 14, web site: http://www.fst-usmba.ac.ma

Abstract

Embedded block coding with optimized truncation (EBCOT) is an important feature of the latest digital still-image compression standard, JPEG2000; however, it consumes more than 50% of the computation time in the compression process. In this paper, we propose a new high speed VLSI implementation of the EBCOT algorithm. The main concept of the proposed architecture is based on parallel access to memories, and uses an efficient design of the context generator block. The proposed architecture is described in VHDL language, verified by simulation and successfully implemented in a Cyclone II and Stratix III FPGA. It provides a major reduction in memory access requirements, as well as a net increase of the processing speed as shown by the simulations.

Keywords: *JPEG20, EBCOT algorithm, VLSI architecture, FPGA implementation.*

1. Introduction

JPEG2000 is the latest still image compression standard, developed by ISO/IEC JTC1/SC29/WG1 (commonly referred to as the Joint Photographic Experts Group JPEG) [1]. JPEG 2000 is not only a competitive compression performance, but also provides many new features for different types of still images [2]. It offers quality scalability, resolution scalability, region of interest (ROI) coding, and supports both lossless and lossy coding in the same framework.

All these features are possible by adoption of the Discrete Wavelet Transform (DWT) and Embedded Block Coding with Optimal Truncation (EBCOT) originally proposed by Taubman [3]. Unfortunately, both algorithms are computation and memory intensive.

The EBCOT is one of the main resources intensive components of JPEG 2000, it accounts for nearly 50% of the total computation time of encoding process [4,6], and then it represents the most critical part in the design and implementation of the JPEG2000 standard. Besides the intensive computation, EBCOT needs massive memory locations. In conventional architectures, the block coder requires at least 20K-bit memory [10]. In order to decrease the EBCOT algorithm time execution two main speedup methods have been suggested. Sample skipping (SS) and Group Of Column Skipping (GOCS) [4]. In the first, one skips no operating samples and in the second all non operating columns are skipped, i.e. all the four bits of the column are skipped. This last technique allows to save a clock cycle naturally wasted in the SS method when a complete column is empty.

Many implementations of hardware architectures have been proposed and designed for EBCOT algorithm to improve the encoding speed, such as Andra's state-machine based bit plane encoder [9], which presents VLSI architecture for embedded bit-plane encoding in JPEG 2000 that reduces the number of memory accesses. This architecture has been implemented in VHDL and the estimated frequency of operation was 200 MHz. Chaing and al. [10] have proposed a Pass-Parallel Context Modeling (PPCM) to implement the EBCOT algorithm, this implementation can work at 180 MHz. They claimed that when compared with the previous context-modeling architectures, there solution improve the throughput rate up to 25%.

In this paper we present an efficient VLSI architecture for EBCOT. It's based on an optimized data organization and a new memory arrangement as well as a simple state machine and combinatorial logics of encoding part. Our proposed architecture makes the four bits to be processed and their neighbors available at one clock cycle and consequently a complete column is processed in only four clock cycles during each pass. This proposed architecture is implemented on FPGA without using any external memory.

The following part of the paper is organized as follows. Section II reviews the EBCOT algorithm; Section III describes the analysis of bit plane coding algorithm. The proposed EBCOT architecture is presented in section IV; the hardware implementation performances and results discussion are described in section V, and section VI concludes the work.

2. EBCOT Algorithm

As illustrated in Fig. 1, the encoding process in JPEG2000 standard follows the typical still image encoding operations. The DWT block transfers the image information from spatial domain to frequency domain and removes the spatial correlation. The redundant information can be rejected by quantization process which is the mainly lossy block in JPEG2000 standard. After the quantization step, many coefficients become zero then the entropy coder can encode the quantized coefficients more efficiently and generates the compressed bit stream. The entropy coding and generation of compressed bit stream in JPEG2000 are two tiers coders, tier-1 is a context based adaptive arithmetic coder and tier-2 is a bit stream layer formation.

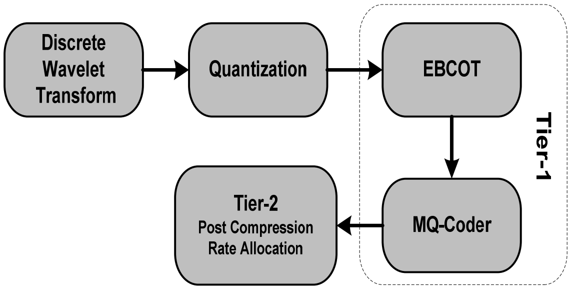


Fig. 1 Functional block diagram of JPEG2000 standard.

2.1Tier-1 coding: context based adaptative arithmetic coder

2.1.1 Bit-plane coding:

The key of JPEG2000 is the EBCOT algorithm. The DWT sub-bands are partitioned into relatively small blocks, called code-blocks (typically 64×64 or 32×32) [1, 2]. The code-blocks are encoded independently. Each code-block is decomposed into n bit-planes. Sequentially, they are encoded from the most significant bit-plane (MSB) to the least significant bit-plane (LSB). Each bit-plane is partitioned into a set of stripe, which spans the full width of the code block and consists of four rows. The stripes are scanned from the top to the bottom one by one. Within each stripe, the scan proceeds column by column, within a column, each sample location is scanned by a top-down manner, until all samples of the column have been visited as shown in Fig. 2.

The encoding process is done by fractional bit-plane coding (BPC) mechanism to create a context and a binary decision value for each bit position. JPEG2000 uses the EBCOT algorithm for the BPC.

This algorithm encodes each generated bit-plane in one of three coding passes [2, 3]: significant propagation pass (PASS-1), magnitude refinement pass (PASS-2), and cleanup pass (PASS-3).. For a bit-plane, these three passes are processed sequentially. The Pass-1 processes coefficients which are insignificant [3] and have at least one significant neighbour among its 8 immediate neighbours, Pass-2 processes all significant coefficients except those becoming significant in Pass-1, finally Pass-3 processes all remaining coefficients not encoded in the Pass-1 and Pass-2. According to the information contained in each bit, four coding primitives are used to generate its context: zero coding (ZC), sign coding (SC), magnitude refinement (MR), and run-length coding (RLC) [3].

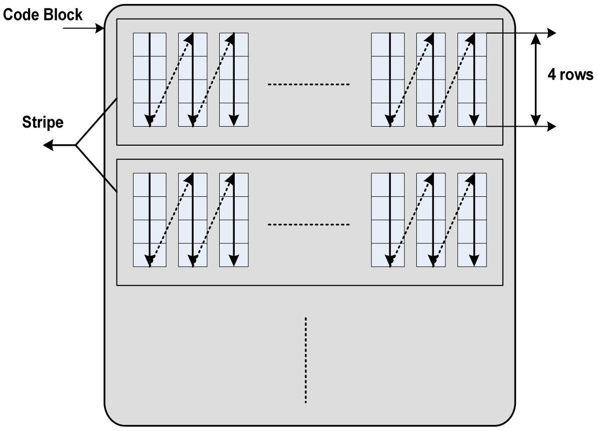


Fig. 2 The scanning order within a bit-plane.

The generated contexts are based on the contextual information (the sign and significance states of the 8-connect neighbors) of the sample scanned in current coding pass. A detailed description about these coding primitives can be found in references [3] and [11].

2.1.2 Arithmetic Coding:

The BPC outputs are entropy encoded using a MQ-coder which is a derivative of the Q-coder [12]. According to the provided context, the coder chooses a probability for the bit to encode, among predetermined probability values supplied by the JPEG 2000 standard and stored in a look-up table. From this probability the MQ-coder progressively generate the compressed code-words. These data are the output of the first tier, and send to Tier 2 for further selection to form the final JPEG2000 bitstream.

2.2 Tier-2 coding: bit stream layer formation

In this second step, each code-block is efficiently represented as a layer and block summary information [2]. A layer is a consecutive bit-plane coding passes from each code-block in a tile, including all sub-bands of all components in the tile. The block summary information consists of the length of compressed code words of the code-block and the truncation point between the bit-stream layers. The compressed code-words generated in the Tier-1 coding step are encoded using a Tag Tree coding mechanism.

3. Analysis of the algorithm

Table 1 shows the complexity estimation for JPEG2000 coding obtained by using the modified software implementation [13]. The run-time table uses a P-IV1.2G CPU, 256M RAM, and VC++6.0 with WINXP system. The size of the test image Lena is 512 x 512 pixels, and we use a configuration with the flowing parameters (lossless and lossy filter, 4 levels wavelet decomposition and one layer with spatial scalability).
Our results (see table 1) as well as others works [5, 6] clearly show that the EBCOT algorithm takes the great part of the processing time compared to the others blocks, this is because EBCOT operations are bit-level processing as illustrated in Fig. 3, which requires important memory resources and several memory accesses.

Table 1: execution time for JPEG2000 encoder using the 512 x 512 Lena image

JPEG2000 blocks	Lossless compression	Lossy compression
DWT	11,4 %	21,2 %
EBCOT	67,8 %	61,3 %
MQ-Coder	18,2 %	14,1 %
Others	2,6 %	3,4 %

Therefore, the key to increase the processing speed of JPEG2000 consists of both new researches and new developments of more efficient VLSI system architecture of EBCOT algorithm.
In order to optimize the hardware design of EBCOT block, a detailed analysis of the algorithm is needed. Fig. 4 shows the analysis results obtained using a 512x512 Lena image which is decomposed by (5,3) filter with 4 levels decomposition for DWT block. Each bit in a bit-plane is encoded only in one of the three coding passes and skipped in the two others.
At first, all samples of the first bit-plane (MSB) are insignificants and encoded in PASS-3 of the coding process, for the lower bit-plane, some samples with neighbouring significant bits will be encoded in Pass-1 and bits which have been significant will be encoded in Pass-2. Therefore from Fig. 4 only a small number of coefficients are encoded by all three coding passes, so two speed-up methods have been proposed in order to accelerate the encoding process: sample skipping (SS) and group-of-column skipping (GOCS) [4, 7, 8]. The key idea of the SS method is to skip those no-operation samples in a single column. The SS is more efficient compared with the straightforward method, but a clock cycle is still wasted when a stripe column is “empty”, that means none of the samples of the stripe column belongs to the current coding pass. Therefore, the second speed-up method, GOCS, is designed to further improve the processing speed. It skips a group of “empty columns” simultaneously at the cost of an extra GOCS memory. Besides, the number of column in a group is a compromise between processing speed and area cost.

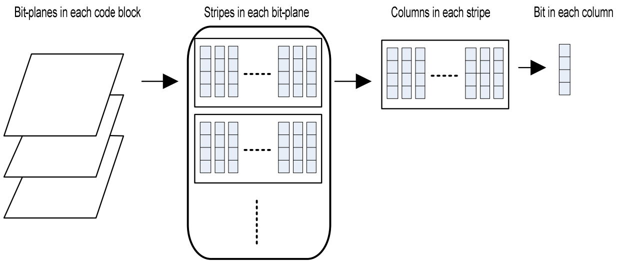


Fig. 3 The hierarchy of a code block.

In this work we adopt a hierarchical scanning for a given code block ordered from lower level (bit) to upper level (bit-plane) as illustrated in Fig. 3. If we can skip from upper level, several iterations will be saved.
The bit-plane coding performs context selection by examining state information for the surrounding neighbours of a sample. Three state variables are necessary for the context formation algorithm (significance state variable, magnitude refinement state variable and visited state variable), as defined in [3] and [11]. These variables are stored in three memories and two others memories are needed to store the sign bit-plane and magnitude bit-plane. Each memory is 4K bits to support the maximum block size. Hence, it is essential to reduce internal memory by using the optimized memory saving algorithm [7].

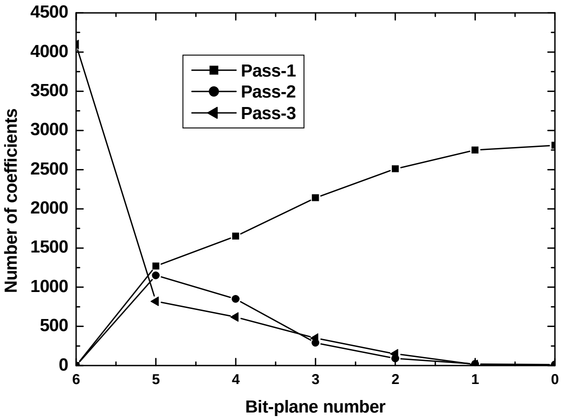


Fig. 4 Coefficients distribution in 3 passes with 64×64 sub-band from 512×512 Lena image.

4. Proposed architecture

In the proposed architecture, a complete column is processed in a single clock-cycle. The four bits to be encoded and their two neighbours are all available at the same time. Therefore to increase the speed of computation and reduce the memory requirement for EBCOT, we exploit the parallel access to memories and propose a new design of context generator for EBCOT tier-1 architecture. The key idea is to bypass the redundant coefficients-bits in each coding pass; it can be done by adopting a new method in the data organisation and memory arrangement. This method proves the efficiency of reducing both access number and memories bandwidth.
The proposed hardware architecture of EBCOT algorithm is presented in Fig. 5. The architecture reads the DWT coefficients data (LL, LH, HL, and HH sub-bands) from the code blocks memories, carry out the discrete wavelet transform, and output the context and data.
This architecture is based on an original register modules (data register module, and state variables register module) communicating through by using two controller blocks and working in parallel with a tight synchronization.
The block diagram of Fig. 5 consists mainly of: 1) Memory blocks 2) Switches and counters 3) Context generator and data selector (mux). All these blocks are controlled and synchronized by two state machines; one manages the pass of the algorithm and the second controls the columns processing in each pass, these blocks are described below.

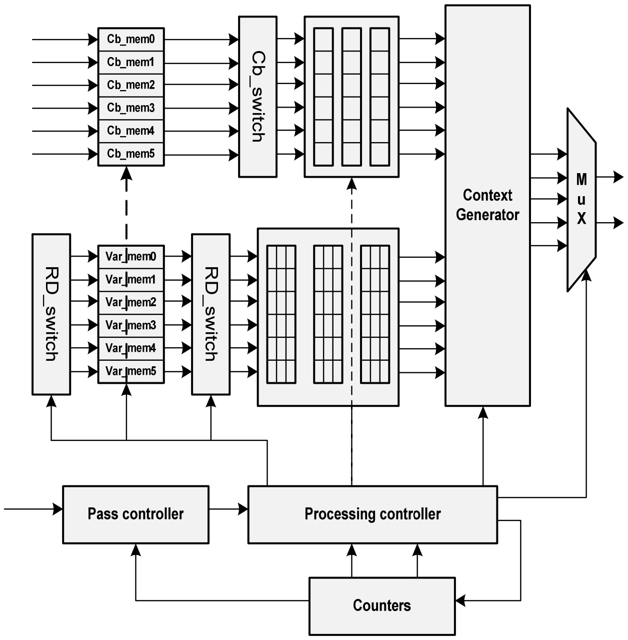


Fig. 5 Block diagram of the proposed EBCOT hardware implementation.

4.1 Data organization and memory arrangement

In order to achieve an efficient data and state variables memories access and to reduce the required memory access clock cycle, we propose a new data arrangement and memory organization to implement the Bit-plane coding. As shown in Fig. 6 memory blocks are organized in six partitions.

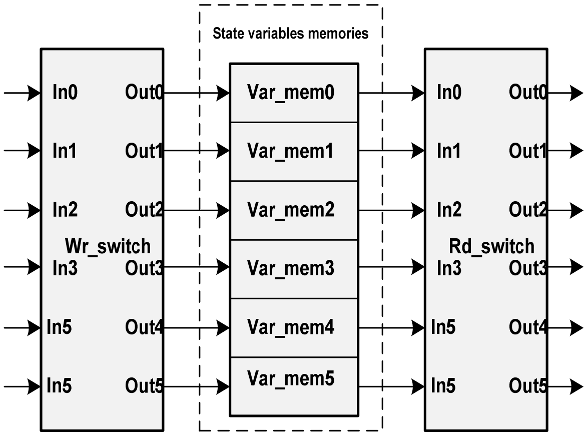


Fig. 6. State variables memories.

MEM0 to MEM5 contain the state variables data. The same organization is adopted for code block coefficients. In a single clock cycle they supply the four bits to be processed and their vertical neighbours at the same time (Fig. 7).
By using this memory arrangement, we can:

- Reduce the complexity of addressing.

- Perform read and write operations at the same clock cycle.
- Prevent the operational conflicts (simultaneous read or write).

Within this arrangement, samples of two nearby stripes, the previous and the next stripes, are loaded into the state variables register simultaneously in only one clock cycle. So, only a single clock cycle is spent by reading data from the six memories and shifting state variable register.

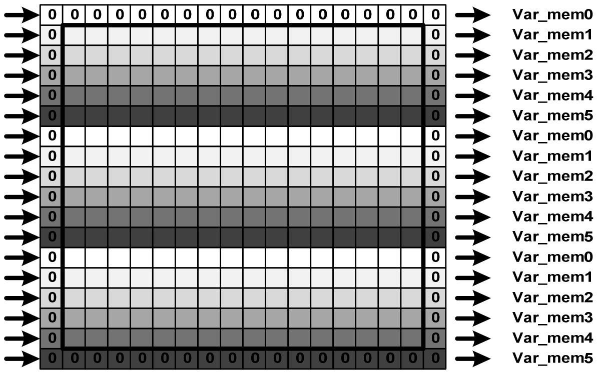


Fig. 7 Code block lines and code block memories association.

For the code block memories, we use a tile splitter block, which extracts the code blocks from each memory of wavelet coefficients sub-band. The extracted data are stored in six memories, in each one; a line of code block is stored as shown in Fig.8. The register blocks are implemented as two six parallel shift registers, one for code block coefficients and the second for state variables. The data shift from memories to registers column by column.

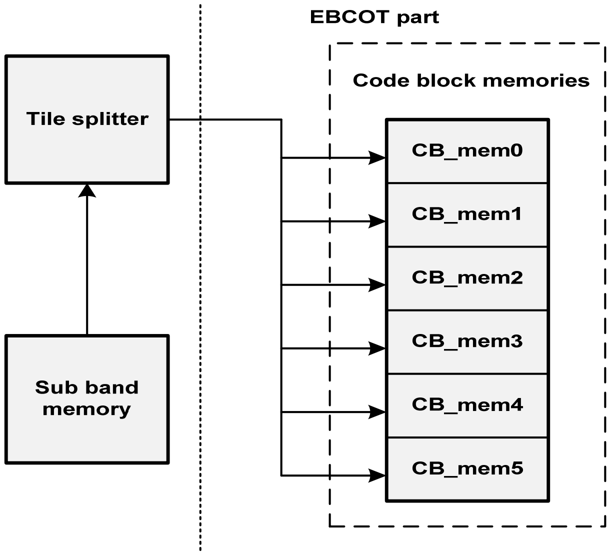


Fig. 8 Code block memories.

4.2 Switches and counter

To manage the code block memories and state variable memories outputs, three switches are used according to the stripe witch is going to be processed. Therefore, there are four switching modes for each switch block.
In this architecture we use one counter to count the columns and stripes of the code block to be processed, and generate the current pass bit plane to be encoded. This counter also allows detecting the position of the bit to be processed inside the code block.

4.3 Context generator and mux

The multiplexer chooses the context from the outputs of ZC(Zero Coding) context block, SC(Sign Coding) context block, MR(Magnitude Refinement) context block, or the hard encoded RLC(Run Length Coding) contexts (17 or 18 contexts).

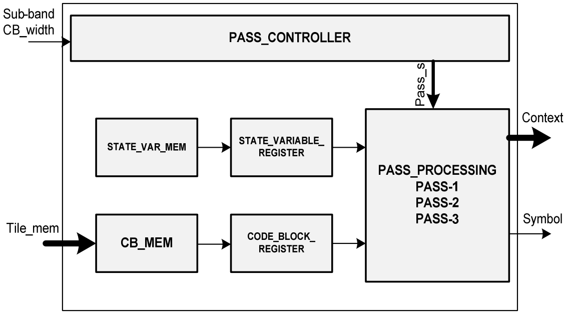


Fig. 9 Context information generator.

For the three passes of EBCOT algorithm, we propose a code block processing model shown in Fig. 9. In this architecture all passes are merged into one component. The pass to be processed is selected by the “pass” signal. This block is managed by the pass-controller block, which is based only on five states, and the symbol is encoded in 3 clock cycles. To better improve the speed of the proposed EBCOT architecture, we use the by pass mode presented in [1].

5. Implementation and experimental results

The proposed architecture was implemented in VHDL. The EBCOT algorithm was also developed using C language (ISO/IEC 15444-1 [1] compatibility) to validate the architecture by comparison with the hardware behavioral.
When synthesized and target to an FPGA ALTERA Cyclone II and Stratix III [14, 15], our design performs at 285 MHz.
Table 2 gives the complete device usage summary. We can see that our design spent 1.5K gates and 40K bits of internal memory.

5.1 Implementation Results

The proposed EBCOT architecture is tested by encoding three test images: Baboon, Jet and Lena with size of 512x512. The code block size is 64×64 or 32×32, and each coefficient can be represented by 12 bits of width.

Table 2: synthesis results of EBCOT block implemented in a altera Cyclone II and stratix III.

	Cyclone II	**Stratix III**
Total logic elements	1,501 (3%)	1,501 (1%)
Total registers	365	234
Total memory bits	40196 (11%)	40196 (5%)
CLK(EBCOT)	225 MHZ	285 MHZ
Operating voltage	1.8 V	1.8 V

In table 3 we show a comparison between our results and those of the most recent work [4, 16]. It is clearly shown that our proposed architecture reduces the processing time by about 45%. This decrease in the processing time are mainly explained by the saving both the wasted clock cycle in the SS method and the required additional clock cycles for memory access in the GOCS architecture.

Table 3: performance of proposed architecture compared with other techniques.

Architecture	**Area (cells)**	**Cycles /code-block**	**CLK (MHz)**
Single Sample[8]	631	156590	51.7
Sample-Skipping[14]	710	89170	38.6
This work	985	41250	285

A high processing frequency is achieved with a suitable number of the cycles/codes-block, thus the proposed architecture is faster than SS [4] and GOCS [16] methods, with low number of the required clock cycles, it reduces the processing time by about 45% as shown in Fig. 10. This increasing in the speed of the EBCOT algorithm is mainly due to the minimization of the number of memory access by adopting an efficient memory architecture to store the state variables and the code block data. The proposed architecture not only overcomes the complexity of state machine, but also has faster computation than PPCM.

However the proposed architecture needs some additional hardware resources, which is relatively low compared to the gain in the processing time, and represents a good compromise between speed and hardware resources for applications such as digital cinema.

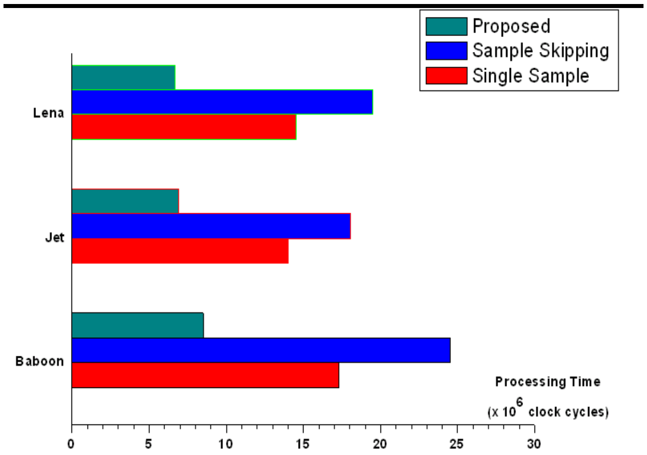


Fig. 10 Comparison of proposed architecture versus other techniques.

5.2 Integration

The proposed EBCOT architecture was first integrated in the JPEG 2000 encoder. It is designed for I-frames in standard definition television (SD, 720 x 480 30 fps) format at 54 MHz and supports high-definition television (HD720p, 1280 x 720 30 fps) format at 100 MHz. Our EBCOT architecture is capable of processing Motion JPEG2000 in real time with 12 bits Bitdepth, 4:4:4 video encoding and a compression ratio of 11.

6. Conclusion

We have proposed an efficient VLSI architecture to implement the Bit-plane coding. The design was implemented in VHDL, and synthesized and routed in an ALTERA Cyclone II and Stratix III FPGA.

This new architecture is based on a parallel access to memories, and uses a new design of context generator block. A working frequency of 285 MHz is achieved, and the number of the required clock cycles is reduced, which increase the processing speed, by comparison with previous works.

The proposed EBCOT encoder is fully compatible with ISO/IEC 15444-1 [1], and can be adopted for supporting real-time applications rates.

The system is secure because no external memory is used and the data flow is protected during the whole encoding process. So, it can be widely used in the application of the futur-generation digital cinema.

References

[1] ISO/IEC 15444-1: Information Technology-JPEG 2000 image coding system-Part 1: Core coding system, 2000.

[2] D. S. Taubman and M. W. Marcellin, JPEG2000: Image Compression Fundamentals, Standards and Practice. Norwell, MA: Kluwer, 2002.

[3] D. Taubman, “High Performance Scalable Image Compression with EBCOT”, IEEE Transactions on Image Processing, Vol. 9, No. 7, July 2000, pp. 1158-1170.

[4] M. D. Adams and F. Kossentini, “Jasper: a software-based JPEG-2000 codec implementation,” Proc. IEEE Int. Conf. Image Processing, vol. 2, pp. 53-56, Sep. 2000.

[5] M. Rabbani, and R. Joshi, “An Overview of the JPEG2000 Still Image Compression Standard”, Signal Processing: Image Communication Journal, Vol. 17, No. 1, October 2001.

[6] M. Dyer, D. Taubman, and S. Nooshabadi, “improved throughput arithmetic coder for JPEG 2000,” accepted in IEEE Int, Conf. Image Processing, pp.2817-2820, 2004.

[7] K.-F. Chen, C.-J. Lian, H.-H. Chen, and L.-G. Chen, “Analysis and architecture design of EBCOT for JPEG2000,” Proc. IEEE Int. Symp. Circuits and Systems, vol. 2, pp. 765-768, May 2001.

[8] C.-J. Lian, K.-F. Chen, H.-H. Chen, and L.-G. Chen, “ Analysis and architecture design of block-coding engine for EBCOT in JPEG 2000,” IEEE Trans. Circuits and Systems for Video Technology, vol. 13, pp. 219-230, March 2003.

[9] Paul R. Schumacher, “An Efficient JPEG2000 Tier-1 Coder Hardware Implementation for Real-Time Video Processing”, IEEE Transactions on Consumer Electronics,Vol. 49, No. 4, November 2003.

[10] Kishore Andra, Chaitali Chakrabarti and Tinku Acharya, “A High Performance JPEG2000 Architecture”, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 13, No. 3, pp 209-218, March 2003.

[11] ISO/IEC JTC 1/SC 29/WG 1 WG1N1878, JPEG 2000 Verification Model 8.5 (Technical description), September 13, 2000.

[12] Cyclone-II platform FPGAs: Complete Data Sheet. ALTERA. [Online]. Available: http://www.altera.com.

[13] Stratix-III platform FPGAs: Complete Data Sheet. ALTERA. [Online]. Available: http://www.altera.com.

[14] Y. Li, R.E. Aly, B. Wilson and M.A. Bayoumi, “Analysis and enhancements for EBCOT in high speed JPEG2000 architectures,” Mid-west Symp. on Ckts. And Systems, vol.2, pp.207-210, Aug. 2002.

[15] Yun-Tai Hsiao, Hung-Der Lin, Kun-Bin Lee and Chein-Wei Jen, “High-Speed Memory-Saving Architecture for the Embedded Block Coding in JPEG2000”, IEEE International Symposium on Circuits and Systems, Vol. 5, pp 133-136, May 2002.

Anass MANSOURI received M.S. and Ph.D degrees in Microelectronics and Telecommunication from Faculty of sciences & technology, Fes, MOROCCO, in 2005 and 2009, respectively. He is a Assistant Professor in National School of Applied Sciences, Fes.
His major research interests include VLSI and embedded architectures design, video and image Processing.

Ali AHAITOUF received the Ph.D. degrees in electronics from the Metz University in France 1992. He is a Professor in electrical engineering department at Faculty of sciences & techniques, Fes, MOROCCO, when he obtained the Doctor Title in Physics at 1998.
His major research interests include Digital and Analog VLSI architecture, EMC Simulation and Physics of Semiconductor Components. He is managing the Microelectronics and Components research group.

Farid ABDI received the Ph.D. degrees in Physics from the Metz University in France 1992. He is a Professor in electrical engineering department at Faculty of sciences & techniques, Fes, MOROCCO.
His major research interests include Optical Components, Image, Audio and video processing. He is managing the optical research group.

IJCSI CALL FOR PAPERS JANUARY 2012 ISSUE

Volume 9, Issue 1

The topics suggested by this issue can be discussed in term of concepts, surveys, state of the art, research, standards, implementations, running experiments, applications, and industrial case studies. Authors are invited to submit complete unpublished papers, which are not under review in any other conference or journal in the following, but not limited to, topic areas. See authors guide for manuscript preparation and submission guidelines.

Accepted papers will be published online and indexed by Google Scholar, Cornell's University Library, DBLP, ScientificCommons, CiteSeerX, Bielefeld Academic Search Engine (BASE), SCIRUS, EBSCO, ProQuest and more.

Deadline: 30th November 2011
Notification: 04th January 2012
Revision: 12th January 2012
Online Publication: 31st January 2012

- Evolutionary computation
- Industrial systems
- Evolutionary computation
- Autonomic and autonomous systems
- Bio-technologies
- Knowledge data systems
- Mobile and distance education
- Intelligent techniques, logics, and systems
- Knowledge processing
- Information technologies
- Internet and web technologies
- Digital information processing
- Cognitive science and knowledge agent-based systems
- Mobility and multimedia systems
- Systems performance
- Networking and telecommunications
- Software development and deployment
- Knowledge virtualization
- Systems and networks on the chip
- Context-aware systems
- Networking technologies
- Security in network, systems, and applications
- Knowledge for global defense
- Information Systems [IS]
- IPv6 Today - Technology and deployment
- Modeling
- Optimization
- Complexity
- Natural Language Processing
- Speech Synthesis
- Data Mining

For more topics, please see http://www.ijcsi.org/call-for-papers.php

All submitted papers will be judged based on their quality by the technical committee and reviewers. Papers that describe on-going research and experimentation are encouraged.
All paper submissions will be handled electronically and detailed instructions on submission procedure are available on IJCSI website (www.IJCSI.org).

For more information, please visit the journal website (www.IJCSI.org)

© IJCSI PUBLICATION 2011

www.IJCSI.org

IJCSI

The International Journal of Computer Science Issues (IJCSI) is a well-established and notable venue for publishing high quality research papers as recognized by various universities and international professional bodies. IJCSI is a refereed open access international journal for publishing scientific papers in all areas of computer science research. The purpose of establishing IJCSI is to provide assistance in the development of science, fast operative publication and storage of materials and results of scientific researches and representation of the scientific conception of the society.

It also provides a venue for researchers, students and professionals to submit ongoing research and developments in these areas. Authors are encouraged to contribute to the journal by submitting articles that illustrate new research results, projects, surveying works and industrial experiences that describe significant advances in field of computer science.

<u>Indexing of IJCSI</u>

1. Google Scholar
2. Bielefeld Academic Search Engine (BASE)
3. CiteSeerX
4. SCIRUS
5. Docstoc
6. Scribd
7. Cornell's University Library
8. SciRate
9. ScientificCommons
10. DBLP
11. EBSCO
12. ProQuest

© IJCSI PUBLICATION
www.IJCSI.org