



# IJCSI

## **International Journal of Computer Science Issues**

**Volume 9, Issue 1, No 3, January 2012  
ISSN (Online): 1694-0814**

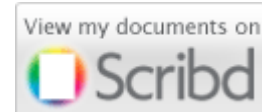
**© IJCSI PUBLICATION  
[www.IJCSI.org](http://www.IJCSI.org)**

**IJCSI proceedings are currently indexed by:**



**Cogprints**

**Google scholar**



**SciRate.com**

**CiteSeer<sup>x</sup> beta**



**Q·Sensei BETA**

**DOAJ** DIRECTORY OF OPEN ACCESS JOURNALS



**ProQuest**

## **IJCSI Publicity Board 2012**

**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 2012**

### **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**

CMSSP Lab, Block A, 2nd Floor, Faculty of Engineering,  
MultiMedia University, Malaysia

**Dr Rajender Singh Chhillar**

Professor  
Department of Computer Science & Applications,  
M. D. University, Haryana, India

## **EDITORIAL**

In this first edition of 2012, 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 3<sup>rd</sup> question in 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 9, Issue 1, No 3, January 2012 (IJCSI Vol. 9, Issue 1, No 3). The acceptance rate for this issue is 30.1%.

IJCSI Editorial Board  
January 2012 Issue  
ISSN (Online): 1694-0814  
© IJCSI Publications  
[www.IJCSI.org](http://www.IJCSI.org)



## **IJCSI Reviewers Committee 2012**

- 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 Mouggiakakou, 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 Iftekhhar 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. Koteswar 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. Shoukat Ullah, Govt. Post Graduate College Bannu, Pakistan
- Dr. Vivian Augustine, Telecom Zimbabwe, Zimbabwe
- Mrs. Mutalli Vatile, 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
- 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. Hafeez Ullah Amin, Kohat University of Science & Technology, Pakistan
- Dr. Professor Vikram Singh, Ch. Devi Lal University, Sirsa (Haryana), 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 Kandeegan, 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 Information 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
- 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
- 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, Ethiopia
- 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 Bilai, India
- Mr. Kuldeep Yadav, IIIT Delhi, India
- Dr. Naoufel Kraiem, Institut Supérieur d'Informatique, Tunisia
- Prof. Frank Ortmeier, Otto-von-Guericke-Universität 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. 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
- 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

- Mr. Ojesanmi Olusegun, Ajayi Crowther University, Oyo, Nigeria
- Ms. Mamta Juneja, RBIEBT, PTU, India
- Prof. Chandra Mohan, John Bosco Engineering College, India
- Dr. Bodhe Shrikant K., College of Engineering, Pandhapur, Maharashtra, INDIA
- Dr. Sherif G. Aly, The American University in Cairo, Egypt
- Mr. Sunil Kashibarao Nayak, Bahirji Smarak Mahavidyalaya, Basmathnagar Dist-Hingoli., India
- Prof. Nikhil gondaliya, G H Patel College of Engg. & Technology, India
- Mr. Nisheeth Joshi, Apaji Institute, Banasthali University, India
- Mr. Nizar, National Engineering School of Monastir, Tunisia
- Prof. R. Jagadeesh Kannan, RMK Engineering College, India
- Prof. Rakesh.L, Vijetha Institute of Technology, Bangalore, India
- Mr B. M. Patil, Indian Institute of Technology, Roorkee, Uttarakhand, India
- Dr. Intisar A. M. Al Sayed, Associate prof./College of Science and IT/Al Isra University, Jordan
- Mr. Thipendra Pal Singh, Sharda University, K.P. III, Greater Noida, Uttar Pradesh, India
- Mrs. Rajalakshmi, JIITU, India
- Mr. Shrikant Ardhapurkar, Indian Institute of Information Techonology, India
- Ms. Hemalatha R, Osmania University, India
- Mr. Hadi Saboohi, University of Malaya - Faculty of Computer Science and Information Technology, Malaysia
- Mr. Sunil Kumar Grandhi, Maris Stella College, India
- Prof. Shishir K. Shandilya, NRI Institute of Science & Technology, INDIA
- Dr. Umesh Kumar Singh, Vikram University, Ujjain, India
- Prof. Prasun Ghosal, Bengal Engineering and Science University, India
- Dr. Nagarajan Velmurugan, SMVEC/Pondicherry University, India
- 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. Palaniyappan, K7 Virus Research Laboratory, India
- Mr. Guanbo Zheng, University of Houston, main campus, USA
- Mr. Arun Kumar Tripathi, Krishna Institute of Engg. and Tech-Ghaziabad, Affiliated to UPTU, India
- Mr. Iqbaldeep Kaur, PTU / RBIEBT, India
- Mr. Amit Choudhary, Maharaja Surajmal Institute, New Delhi, India
- Mrs. Vasudha Bahl, Maharaja Agrasen Institute of Technology, Delhi, India
- Dr. Ashish Avasthi, Uttar Pradesh Technical University, India
- Dr. Manish Kumar, Uttar Pradesh Technical University, India
- Prof. Vinay Uttamrao Kale, P.R.M. Institute of Technology & Research, Badnera, Amravati, Maharashtra, India
- Mr. Suhas J Manangi, Microsoft, India
- Mr. Shyamalendu Kandar, Haldia Institute of Technology, 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
- Mr. Senthilnathan T, Sri Krishna College of Engineering and Technology, 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
- Mrs. Manjula K A, Kannur University, India
- Mrs. Sasikala R., K S R College of Technology, 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
- 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
- Dr. Manoj Wadhwa, Echelon Institute of Technology Faridabad, India
- Mr. Zeashan Hameed Khan, Universit  de Grenoble, France
- Mr. Arup Kumar Pal, Indian School of Mines, Dhanbad, India
- Dr. Pouya, Islamic Azad University, Naein Branch, Iran
- Prof. Jyoti Prakash Singh, Academy of Technology, India
- Mr. Muraleedharan CV, Sree Chitra Tirunal Institute for Medical Sciences & Technology, India
- Dr. E U Okike, University of Ibadan, Nigeria Kampala Int Univ Uganda, Nigeria
- Dr. D. S. Rao, Chitkara University, India
- Mr. Peyman Taher, Oklahoma State University, USA
- Dr. S Srinivasan, PDM College of Engineering, India
- Dr. Rafiqul Zaman Khan, Department of Computer Science, AMU, Aligarh, India
- Ms. Meenakshi Kalia, Shobhit University, India
- Mr. Muhammad Zakarya, Abdul Wali Khan University, Mardan, Pakistan, Pakistan
- Dr. M Gobi, PSG college, India
- Mr. Williamjeet Singh, Chitkara Institute of Engineering and Technology, India
- Mr. G.Jeyakumar, Amrita School of Engineering, India
- Mr. Osama Sohaib, University of Balochistan, Pakistan
- Mr. Jude Hemanth, Karunya University, India
- Mr. Nitin Rakesh, Jaypee University of Information Technology, India
- Mr. Harmunish Taneja, Maharishi Markandeshwar University, Mullana, Ambala, Haryana, India
- Dr. Sin-Ban Ho, Faculty of IT, Multimedia University, Malaysia
- Dr. Mashiur Rahman, Institute for Molecular Science, Japan
- Mrs. Doreen Hephzibah Miriam, Anna University, Chennai, India
- Mr. Kosala Yapa Bandara, Dublin City University, Ireland.
- Mrs. Mitu Dhull, GNKITMS Yamuna Nagar Haryana, India



- Dr. Chitra A.Dhawale, Professor, Symbiosis Institute of Computer Studies and Research, Pune (MS), India
- Dr. Arun Sharma, GB Technical University, Noida, India
- Mr. Naoufel Machta, Faculty of Science of Tunis, Tunisia
- Dr. Utpal Biswas, University of Kalyani, India
- Prof. Parma Nand, IIT Roorkee, India
- Prof. Mahesh P K, Jnana Vikas Institute of Tevhnology, Bangalore, India
- Dr. D.I. George Amalarethnam, Jamal Mohamed College, Bharathidasan University, India
- Mr. Ishtiaq ahmad, University of Engineering & Technology, Taxila, Pakistan
- Mrs. B.Sharmila, Sri Ramakrishna Engineering College, Coimbatore Anna University Coimbatore, India
- Dr. Muhammad Wasif Nisar, COMSATS Institue of Information Technology, Pakistan
- Mr. Prabu Dorairaj, EMC Corporation, India/USA
- Mr. Neetesh Gupta, Technocrats Inst. of Technology, Bhopal, India
- Dr. Ola Osunkoya, PRGX, USA
- Ms. A. Lavanya, Manipal University, Karnataka, India
- Dr. Jalal Laassiri, MIA-Laboratory, Faculty of Sciences Rabat, Morocco
- Mr. Ganesan, Sri Venkateswara college of Engineering and Technology, Thiruvallur, India
- Mr. V.Ramakrishnan, Sri Venkateswara college of Engineering and Technology, Thiruvallur, India
- Prof. Vuda Sreenivasarao, St. Mary's college of Engg & Tech, India
- Prof. Ashutosh Kumar Dubey, Assistant Professor, India
- Dr. R.Ramesh, Anna University, India
- Mr. Ali Khadair HMood, University of Malaya, Malaysia
- Dr. Vimal Mishra, U.P. Technical Education, India
- Mr. Ranjit Singh, Apeejay Institute of Management, Jalandhar, India
- Mrs. D.Suganyadevi, SNR SONS College (Autonomous), India
- Mr. Prasad S.Halgaonkar, MIT, Pune University, India
- Mr. Vijay Kumar, College of Engg. and Technology, IFTM, Moradabad(U.P), India
- Mr. Mehran Parchebafieh, Douran, Iran
- Mr. Anand Sharma, MITS, Lakshmanagarh, Sikar (Rajasthan), India
- Mr. Amit Kumar, Jaypee University of Engineering and Technology, India
- Prof. B.L.Shivakumar, SNR Sons College, Coimbatore, India
- Mr. Mohammed Imran, JMI, India
- Dr. R Bremananth, School of EEE, Information Engineering (Div.), Nanyang Technological University, Singapore
- Prof. Vasavi Bande, Computer Science and Engineering, Hyderabad Institute of Technology and Management, India
- Dr. S.R.Balasundaram, National Institute of Technology, India
- Dr. Prasart Nuangchalerm, Mahasarakham University, Thailand
- Dr. M Ayoub Khan, C-DAC, Ministry of Communications & IT., 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
- Mr. Amin Daneshmand Malayeri, Young Researchers Club, Islamic AZAD University, Malayer Branch, Iran
- Dr. Arti Arya, PES School of Engineering, Bangalore (under VTU, Belgaum, Karnataka), India

- Mr. Pawan Jindal, J.U.E.T. Guna, M.P., India
- Dr. Soumen Mukherjee, RCC Institute of Information Technology, India
- Dr. Hamid Mcheick, University of Qubec at Chicoutimi, Canada
- Dr. Mokhled AlTarawneh, PhD computer engineering/ Faculty of engineerin/ mutah university, jordan
- Prof. Santhosh.P.Mathew, Saintgits College of Engineering, Kottayam, India
- Ms. Suman Lata, Rayat Bahara institue of engg. & Nanotechnology,Hoshiarpur, India
- Dr. Shaikh Abdul Hannan, Vivekanand College, Aurangabad, India
- Prof. PN Kumar, Amrita Vishwa Vidyapeetham, India
- Dr. P. K. Suri, Department of Computer Science & Applications, Kurukshetra University, Kurukshetra, India
- Dr. Syed Akhter Hossain, Daffodil International University, Bangladesh
- Mr. Sunil, Vignan College, India
- Mr. Ajit Singh, TIT&S Bhiwani, Haryana, India
- Mr. Nasim Qaisar, Federal Urdu Univetrstity of Arts , Science and Technology, Pakistan
- Ms. Rshma, Maharishi Markandeshwar University, India
- Mr. Gaurav Kumar Leekha, M.M.University, Solan (Himachal Pardesh), India
- Mr. Ordinor Tucker, Ministry of Finance Jamaica, Jamaica
- 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
- Dr. G R Sinha, SSCET, India
- Mr. Chetan Sharma, TechMahindra India Ltd., India
- Dr. Nabil Mohammed Ali Munassar, University of Science and Technology, Yemen
- Prof. T Venkat Narayana Rao, Department of CSE, Hyderabad Institute of Technology and Management , India
- Prof. Vasavi Bande, HITAM, Engineering College, India
- Prof. S.P.Setty, Andhra University, India
- Dr. C. Kiran Mai, J.N.T.University,Hyderabad/VNR Vignana Jyothi Institute of Engineering & Technology/, India
- Ms. Bindiya Ahuja, Manav Rachna International University, India
- Mrs. Deepa Bura, Manav Rachna International University, 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
- Mr. Abhishek Shukla, RKGIT, India
- Prof. Manpreet Singh, M. M. Engg. College, M. M. University, Haryana, India
- Mr. Syed Imran, University College Cork, Ireland
- Dr. Intisar Al Said, Associate Prof/Al Isra University, Jordan
- Dr. Namfon Assawamekin, University of the Thai Chamber of Commerce, Thailand
- Dr. Shiv KUMar, Technocrat Institute of Technology-Bhopal (M.P.), India
- Dr. Shahaboddin Shamshirband, Islamic Azad University, Iran
- 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
- Prof. Jakimi, Faculty of Science and technology my ismail University, Morocco
- Dr. R. Manicka Chezian, N G M College, Pollachi - 642 001, Tamilnadu, India
- Dr. P.Dananjayan, Pondicherry Engineering College, India
- Mr. Manik Sharma, Sewa Devi SD College Tarn Taran, 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 Institute of Engineering & Technology, India
- Dr. Anuraag Awasthi, JV Womens University, Jaipur, India
- Dr. Mathura Prasad Thapliyal, Department of Computer Science, HNB Garhwal University (Central 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. 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 Engineering, (Affiliated 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 Pavanam, Adhiparasakthi Engineering College, Melmaruvathur, India
- Dr. Panagiotis Michailidis, University of Western Macedonia, Greece
- Mr. Amir Seyed Danesh, University of Malaya, Malaysia
- Dr. Nadeem Mahmood, Department of computer science, university of Karachi, Pakistan
- 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
- Dr. Hanumanthappa. J, DoS in Computer Science, India
- 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
- Ms. Suneet Kumar, Uttarakhand Technical University/Dehradun Institute of Technology, Dehradun, Uttarakhand, India
- Mr. Rupesh Nasre., Indian Institute of Science, Bangalore., India.
- Mrs. Dimpi Srivastava, Dept of Computer science, Information Technology and Computer Application, MIET, Meerut, India
- Dr. Eva Volna, University of Ostrava, Czech Republic
- 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
- Prof. Asha Ambhaikar, Rungta College of Engineering & Technology, Bilai, India
- Dr. Saurabh Dutta, Dr. B. C. Roy Engineering College, Durgapur, India
- Dr. Mokhled Altarawneh, Mutah University, Jordan
- Mr. Anand Nayyar, KCL Institute of Management and Technology, Jalandhar, India
- Mr. S. A. Ahsan Rajon, Computer Science and Engineering Discipline, Khulna University, Bangladesh
- Ms. Rezarta Mersini, University of Durres, Albania
- Mrs. Deepika Joshi, Jaipuria Institute of Management Studies, India
- Dr. Niraj Shakhakarmi, Prairie View A&M University, (Texas A&M University System), USA
- Mrs. A. Valarmathi, Anna University, Trichy, India
- Dr. K. Balamurugan, Institute of Road and Transport Technology, India
- Prof. K. S. Sridharan, Sri Sathya Sai Institute of Higher Learning, India
- Mr. Okumoku-Evroro Oniovosa, Delta State University, Abraka, Nigeria
- Mr. Rajiv Chopra, GTBIT, Delhi, India
- Mr. Harish Garg, Department of Mathematics, IIT Roorkee, India
- Mr. Ganesh Davanam, Sree Vidyanikethan Engineering College, India
- Mr. Bhavesh Shah, VIT, India
- Dr. Suresh Kumar Bhardwaj, Manav Rachna International University, India
- Dr. Muhammad Nawaz Khan, School of Electrical Engineering & Computer Science, Pakistan
- Ms. Saranya, Bharathidasan University, India
- Mr. Sumit Joshi, GRD-IMT, Dehradun, India
- Dr. Mohammed M. Abu Shquier, Tabuk University, School of Computers and Information Technology, Kingdom of Saudi Arabia
- Ms. Shalini Ramanathan, PSG College of Technology, India
- Mr. S. Munisankaraiyah, Geethanjali College of Engineering & Technology, Hyderabad, India
- Dr. Satyanarayana, KL University, India
- Mr. Sarin CR, Anna University, India
- Mr. Sayed Shoaib Anwar, Mahatma Gandhi Mission College of Engineering, India
- Mrs. Gunjan, JSSATE, Noida, India
- Dr. Ramachandra V. Pujeri, Anna University, India
- Mrs. Antima Singh Puniya, Shobhit University, Meerut, India
- Dr. Avdhesh Gupta, College of Engineering Roorkee, India
- Ms. Shiva Prakash, Madan Mohan Malaviya Engg. College, Gorakhpur, India
- Dr. Kristijan Kuk, School of Electrical Engineering and Computer Science Applied Studies, Belgrade, Serbia
- Prof. Dinesh Vitthalrao Rojarkar, Govt. College of Engineering, Chandrapur, India
- Prof. Lalji Prasad, RGTU/TCET, Indore, India
- Dr. A. John Sanjeev Kumar, Thiagarajar College of Engineering, Madurai, Tamilnadu, India
- Mr. Harishbabu Kalidasu, Priyadarshini Institute of Technology and Science, Tenali, Guntur(DT), Andhra Pradesh, India
- Prof. Vaitheeshwaran, Priyadarshini Indira Gandhi College of Engineering, India
- Mrs. P. Salini, Pondicherry Engineering College, India

- Mr. Vivek Bhambri, Desh Bhagat Institute of Management and Computer Sciences, Mandi Gobindgarh(Punjab), India
- Mr. Slavko Zitnik, Faculty of Computer and Information Science Ljubljana, Slovenia
- Ms. Sreenivasa Rao, CMJ University/Yodlee Infotech, India

## **TABLE OF CONTENTS**

<b>1. Solving Problems in Software Applications through Data Synchronization in Case of Absence of the Network</b> <b>Isak Shabani, Betim Cico and Agni Dika</b>	<b>10-16</b>
<b>2. Inversion of Web Service Invocation using Publish/Subscribe Push-Based Architecture</b> <b>Thanisa Numnonda and Rattakorn Poonsuph</b>	<b>17-25</b>
<b>3. Design of a Conceptual Reference Framework for Reusable Software Components based on Context Level</b> <b>V. Subedha and S. Sridhar</b>	<b>26-31</b>
<b>4. Integrated three-dimensional reconstruction using reflectance fields</b> <b>Maria-Luisa Rosas and Miguel-Octavio Arias</b>	<b>32-36</b>
<b>5. Reasoning in Graph-based Clausal Form Logic</b> <b>Alena Lukasova, Martin Zacek and Marek Vajgl</b>	<b>37-43</b>
<b>6. Extraction of Facial Feature Points Using Cumulative Histogram</b> <b>Sushil Kumar Paul, Mohammad Shorif Uddin and Saida Bouakaz</b>	<b>44-51</b>
<b>7. Validity Index and number of clusters</b> <b>Mohamed Fadhel Saad and Adel M. Alimi</b>	<b>52-57</b>
<b>8. An Accurate Arabic Root-Based Lemmatizer for Information Retrieval Purposes</b> <b>Tarek El-Shishtawy and Fatma El-Ghannam</b>	<b>58-66</b>
<b>9. Synthesis of Quantum Multiplexer Circuits</b> <b>Arijit Roy, Dibyendu Chatterjee and Subhasis Pal</b>	<b>67-74</b>
<b>10. Multi-Objective Evolutionary Computation Solution for Chocolate Production System Using Pareto Method</b> <b>Alaa Sheta, Abdel karim Baareh, Mohamed Ababna and Noor Khrisat</b>	<b>75-83</b>
<b>11. Improve and Compact Population in XCSFCA using Polynomial Equation</b> <b>Saeid Goodarzian, Ali Hamzeh and Sattar Hashemi</b>	<b>84-95</b>
<b>12. High-Performance Low-Power Digital Linear Interpolation Filter</b> <b>Magdy El-Moursy and Ahmed G. Abdellatif</b>	<b>96-100</b>
<b>13. Mobile Agent Based Hierarchical Intrusion Detection System in Wireless Sensor Networks</b> <b>Surraya Khanum, Muhammad Usman and Alaa Alwabel</b>	<b>101-108</b>
<b>14. New method to parse invoice as a type the document</b> <b>Mohammed Moujabbir and Mohammed Ramdani</b>	<b>109-114</b>
<b>15. Techniques, Advantages and Problems of Agent Based Modeling for Traffic Simulation</b> <b>Ali Bazghandi</b>	<b>115-119</b>
<b>16. Optimization Parameters of tool life Model Using the Taguchi Approach and Response Surface Methodology</b> <b>Kompan Chomsamutr and Somkiat Jongprasithporn</b>	<b>120-125</b>

<b>17. Temperature effects on the Drain Current in GaN Dual-Gate MESFET using Two-Dimensional Device Simulation</b> <b>Hamida Djelti, Mohammed Feham, Achour Ouslimani and Abed-Elhak Kasbari</b>	<b>126-129</b>
<b>18. Improving Multi agent Systems Based on Reinforcement Learning and Case Base Reasoning</b> <b>Sara Esfandiari, Behrooz Masoumi, Mohammadreza Meybodi and Abdolkarim Niazi</b>	<b>130-138</b>
<b>19. Peer to Peer Networks Management Survey</b> <b>Mourad Amad, Ahmed Meddahi and Djamil Aissani</b>	<b>139-148</b>
<b>20. Human Iris Segmentation for Iris Recognition in Unconstrained Environments</b> <b>Mahmoud Mahlouji and Ali Noruzi</b>	<b>149-155</b>
<b>21. Proposition of Model for CSIRT: Case Study of Telecommunication Company in a Province of Iran</b> <b>Ali Naseri and Omid Azmoon</b>	<b>156-160</b>
<b>22. Protein sequence for clustering DNA based on Artificial Neural Networks</b> <b>Gamal. F. Elhadi, R. M. Farouk and Abdalhakeem. T. Issa</b>	<b>161-167</b>
<b>23. Reliable and Efficient Routing Using Adaptive Genetic Algorithm in Packet Switched Networks</b> <b>Rakesh Kumar and Mahesh kumar</b>	<b>168-174</b>
<b>24. Serious Security Weakness in RSA Cryptosystem</b> <b>Majid Bakhtiari and Mohd Aizaini Maarof</b>	<b>175-178</b>
<b>25. Two Automated Mechanisms to Create Electures and to Videotape Regular Lectures</b> <b>Nael Hirzallah</b>	<b>179-186</b>
<b>26. Rigorous Description Of Design Components Functionality: An Approach Based Contract</b> <b>Abdelhafid Zitouni</b>	<b>187-196</b>
<b>27. Method of operational diagnostic state of flow and calculation of calibration Coefficients using artificial neural networks</b> <b>Safarini Osama</b>	<b>197-200</b>
<b>28. USABILITY of Collaborative Web Surfing Systems in e-Research</b> <b>Akhtar Ali Jalbani, Aneela Yasmin, Gordhan Das Menghwar and Mukhtiar Memon</b>	<b>201-205</b>
<b>29. Search engine optimization with Google</b> <b>Vinit Kumar Gunjan</b>	<b>206-214</b>
<b>30. DSP Implementation of a Power Factor Correction Strategy for BLDC Motor Drive</b> <b>R. Vijayarajeswaran</b>	<b>215-220</b>
<b>31. Modified Secret Sharing over a Single Path in VoIP with Reliable Data Delivery</b> <b>K.Maheswari and M. Punithavalli</b>	<b>221-226</b>
<b>32. Spoken Word Recognition Strategy for Tamil Language</b> <b>AN. Sigappi and S. Palanivel</b>	<b>227-233</b>
<b>33. Implementation of Genetic Algorithm in Predicting Diabetes</b> <b>S.Sapna, A. Tamilarasi and M.Pravin Kumar</b>	<b>234-240</b>
<b>34. Influence of Side Effect of EBG Structures on the Far-Field Pattern of Patch Antennas</b> <b>Fethi Benikhlef and Nourdinne Boukli Hacem</b>	<b>241-245</b>



<b>35. Intelligent Video Object Classification Scheme using Offline Feature Extraction and Machine Learning based Approach</b> <b>Chandra Mani Sharma, Alok Kumar Singh Kushwaha, Rakesh Roshan, Rabins Porwal and Ashish Khare</b>	<b>247-256</b>
<b>36. An Extensible and Secure Framework for Distributed Applications</b> <b>Aneesha Sharma and Shilpi Gupta</b>	<b>257-265</b>
<b>37. A SLA-Aware Scheduling Architecture in Grid System Using Learning Techniques</b> <b>Seyedeh Yasaman Rashida and Amir Masoud Rahmani</b>	<b>266-275</b>
<b>38. Mobility Metrics Estimation and Categorization for SNET Protocols</b> <b>Anita Sethi, Anita Sethi, J. P. Saini, J. P. Saini, Shailendra Mishra and Shailendra Mishra</b>	<b>276-282</b>
<b>39. Using Petri Nets For Resource Management Modeling In The Operating Systems</b> <b>Adalat Karimov and Shahram Moharrami</b>	<b>283-288</b>
<b>40. MPLS - A Choice of Signaling Protocol</b> <b>Muhammad Asif, Zahid Farid, Muhammad Lal and Junaid Qayyum</b>	<b>289-295</b>
<b>41. Block Based Video Watermarking Scheme Using Wavelet Transform and Principle Component Analysis</b> <b>Nisreen I. Yassin, Nancy M. Salem and Mohamed I. El Adaway</b>	<b>296-301</b>
<b>42. Tools for decision support in planning academic needs of actors</b> <b>Oubedda Latifa, Erraha Brahim and Khalfaoui Mohamed</b>	<b>302-306</b>
<b>43. Outlier Detection: Applications And Techniques</b> <b>Karanjit Singh and Shuchita Upadhyaya</b>	<b>307-323</b>
<b>44. Pair of Iris Recognition for Personal Identification Using Artificial Neural Networks</b> <b>K.Saminathan, M.Chithra Devi and T.Chakravarthy</b>	<b>324-327</b>
<b>45. A New Approach to the Data Aggregation in Wireless Sensor Networks</b> <b>Chamran Asgari and Javad Akbari Torkestani</b>	<b>328-335</b>
<b>46. A 100 mA Low Voltage Linear Regulators for Systems on Chip Applications Using 0.18 um CMOS Technology</b> <b>Krit Salah-ddine, Zared Kamal, Qjidaa Hassan and Zouak Mohcine</b>	<b>336-342</b>
<b>47. Automated Text Summarization Base on Lexicales Chain and graph Using of WordNet and Wikipedia Knowledge Base</b> <b>Mohsen Pourvali and Mohammad Saniee Abadeh</b>	<b>343-349</b>
<b>48. Designing Debugging Models for Object Oriented Systems</b> <b>Sujata Khatri and R. S. Chhillar</b>	<b>350-357</b>
<b>49. Designing of RF Single Balanced Mixer with a 65 nm CMOS Technology Dedicated to Low Power Consumption Wireless Applications</b> <b>Raja Mahmou and Khalid Fatah</b>	<b>358-363</b>
<b>50. Analyzing the Complexity of Java Programs using Object-Oriented Software Metrics</b> <b>Arti Chhikara and R.S.Chhillar</b>	<b>364-372</b>
<b>51. Semantic Malware Detection by Deploying Graph Mining</b> <b>Fatemeh Karbalaie, Ashkan Sami and Mansour Ahmadi</b>	<b>373-379</b>

<b>52. A Multi-agent Approach for Space Occupation Problems</b> <b>Jamila Boussaa, Mohammed Sadgal and Aziz Elfazziki</b>	<b>380-389</b>
<b>53. Tradeoff Analysis of Bit-Error-Rate (BER) in Cognitive Radio Based on Genetic Algorithm</b> <b>Shrikrishan Yadav and Krishna Chandra Roy</b>	<b>390-394</b>
<b>54. Evaluation Strategy for Ranking and Rating of Knowledge Sharing Portal Usability</b> <b>D. Venkata Subramanian and Angelina Geetha</b>	<b>395-400</b>
<b>55. Vehicular Ad-hoc Networks</b> <b>Asad Maqsood and Rehanullah Khan</b>	<b>401-408</b>
<b>56. Video Authentication: Issues and Challenges</b> <b>Saurabh Upadhyay and Sanjay Kumar Singh</b>	<b>409-418</b>
<b>57. Novel Design of A Compact Proximity Coupled Fed Antenna</b> <b>Mahdi Ali, Abdennaceur Kachouri and Mounir Samet</b>	<b>419-426</b>
<b>58. ECG Analysis based on Wavelet Transform and Modulus Maxima</b> <b>Mourad Talbi, Akram Aouinet, Riadh Baazaoui and Adnane Cherif</b>	<b>427-435</b>
<b>59. Presentation an approach for scientific workflow distribution on cloud computing data centers servers to optimization usage of computational resource</b> <b>Ahmad Faraahi, Rohollah Esmaeli Manesh, Ahmad Zareie, Mir Mohsen Pedram and Meysam Khosravi</b>	<b>436-451</b>
<b>60. Comparison of Routing Protocols for Locating moving object in Large Scale Cellular Wireless Sensor Network</b> <b>Ola A. Al-Sonosy, Mohammed A. Hashem and Nagwa Badr</b>	<b>452-464</b>
<b>61. Security Analysis of Routing Protocols in Wireless Sensor Networks</b> <b>Mohammad Sadeghi, Farshad Khosravi, Kayvan Atefi and Mehdi Barati</b>	<b>465-472</b>
<b>62. Skin-color Based Videos Categorization</b> <b>Rehanullah Khan, Asad Maqsood, Zeeshan Khan, Muhammad Ishaq and Arsalan Arif</b>	<b>473-477</b>
<b>63. Network Threat Ratings in Conventional DREAD Model Using Fuzzy Logic</b> <b>Ak.Ashakumar Singh and K.Surchandra Singh</b>	<b>478-484</b>
<b>64. Using JQuery with Snort to Visualize Intrusion</b> <b>Alaa El - Din Riad, Ibrahim Elhenawy, Ahmed Hassan and Nancy Awadallah</b>	<b>486-491</b>
<b>65. Performance Evaluation of AODV and DSR with Varying Pause Time and Speed Time Over TCP and CBR Connections in VANET</b> <b>Bijan Paul, Md.Ibrahim and Md. Abu Naser Bikas</b>	<b>493-504</b>
<b>66. Assessment of Water Quality in Coastal Environments of Mohammedia Applying Responses of Biochemical Biomarkers in the Brown Mussel Perna perna</b> <b>Laila El Jourmi and Abdessamad Amine</b>	<b>505-510</b>

# Solving Problems in Software Applications through Data Synchronization in Case of Absence of the Network

Isak Shabani<sup>1</sup>, Betim Çiço<sup>2</sup> and Agni Dika<sup>3</sup>

<sup>1,3</sup> Faculty of Computer and Electrical Engineering, University of Pristina,  
Pristina, 10000, Kosovo

<sup>2</sup> Computer Engineering Department, Polytechnic University Tirana,  
Tirana, 1001, Albania

## Abstract

In this paper, we have presented an algorithm for data synchronization based on Web Services (WS), which allows software applications to work well on both configurations "Online" and "Offline", in the absence of the network. For this purpose is in use Electronic Student Management System (ESMS) at University of Prishtina (UP) with the appropriate module. Since the use of ESMS, because of a uncertain supply of electricity, disconnecting the network and for other reasons which are not under the control of professional staff that manages the performance of this system, has interruption to the online work. In order to continue work in such conditions, are founded adequate solutions to work in offline mode and later data synchronization in normal conditions.

**Keywords:** *Web application, Web services, Data Synchronization, Offline mode.*

## 1. Introduction

With the project of digitalization of academic and administrative affairs in UP, the phase of complete computerized services has begun. For the implementation of this project from staff of Information Technology (IT), in cooperation with the management of UP the required environment of hardware and software is being prepared, respective academic network, data center and software applications are being completed. The results obtained from this work are tested and successfully implemented in the Electronic Student Management System (ESMS) also developed from UP.

The aim of this paper is to provide new results for data synchronization in different platforms through Web services, which allow software applications to run or to be executed online and offline as well [3]. The use of software applications is more productive especially when these applications can work anytime and anywhere without inconveniences related to error messages referred

to the user, such as: "No access to network" or "No access to Server". In this paper, "Offline mode" development for the UP ESMS System, in order to increase confidence in using software applications is presented. However, the challenges for building and managing reliable synchronization algorithms are a major concern and potentially dangerous. Algorithm design for data synchronization through Web Services [7], used to build software applications, which besides the online mode of operation are able to work also in offline mode, in the absence of Internet and power problems, and make these interruptions unnoticed by end users, is considered [5].

## 2. The current state of functioning of systems in Offline mode

Offline systems so far have worked based on several techniques. Among the most advanced techniques is MS Synchronize Framework or shortly MS Sync Framework, which is a platform from Microsoft for synchronizing data from many units and offline Web applications [9]. Sync Framework can be used to access data offline. Synchronization can be used for one-to-one or one-to-many units, which function offline [11]. The synchronization service is incorporated in VS2008 and VS2010 .

This option provides synchronization of data synchronization, file synchronization and synchronization of news publications [2]. This synchronization is summarized in the following services: Sync Services for ADO.NET, Sync Services for File Systems and Sync Services for FeedSync. Thus, SyncServices for ADO.NET enables data synchronization for ADO.NET. SyncServices Service for File Systems, enables synchronization of files between two or more units with the central unit and the SyncServices for FeedSync, enables synchronization of

information in the form of RSS, Atom and distributed transaction [1]. Web applications are preferred over desktop applications because they are available 24x7 and are accessible from anywhere in the network and cloud computing [6]. A couple of years back, the term network implied wired-network but more recently, due to availability of high-speed wireless networks and handy mobile devices, the dependency over wired-network has diminished greatly. Via wireless connectivity, Web applications are now accessible literally from anywhere.

### 3. Data Synchronization

Optimistic replication strategies are attractive in a growing range of settings where weak consistency guarantees can be accepted in return for higher availability and the ability to update data while disconnected. These uncoordinated updates must later be synchronized (or reconciled) by automatically combining non conflict updates, while detecting and reporting conflict updates [8]. The ability to support mobile and remote workers is becoming more and more important for organizations every day. It is critical that organizations ensure users have access to the same information they have when they are in the office. In most cases, these workers will have some sort of laptop, office desktop, Smartphone, or PDA. From these devices, users may be able to access their data directly through VPN connections, Web Servers, or some other connectivity method into the corporate networks [5]. Synchronization gained great importance in modern applications and allows mobility in the context of information technology. Users are not limited to one computer any more, but can take their data with them on a laptop.

#### 3.1 Synchronizing complex objects

Process of synchronizing complex objects will be explained following the scheme as shown in Figure 1.

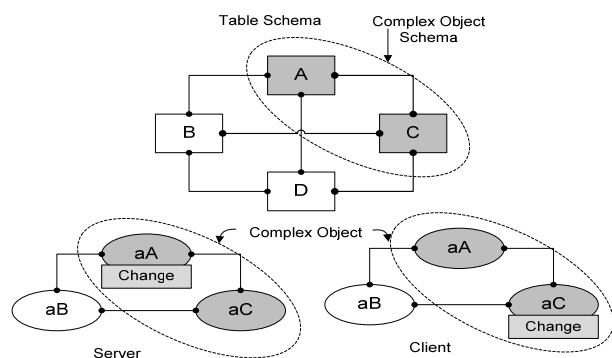


Fig. 1 Synchronization concept for complex objects .

Table A could hold zip code and city address. Table C includes street address and street number. Objects conceptually belong to each other and form a complex address of an object. Objects that conceptually belong to each other in sense of modeling in a real world are treated as a single complex object, thus the conflicts could be detected and if the data in different tables are changed. Normal process of replication would not detect the conflict and merge the data without showing them to the user. Developers can determine the objects of the class which form complex object's class. This information then is used in the time of synchronization. There are three reasons that the synchronization of complex objects is a better process: data model can be normalized and dependent data should not be considered, data model can change, however some conflicts are not detected and conflict detection in object instead of data tables is more understandable for the user.

#### 3.2 Synchronization conflicts

A conflict can only occur when two databases have a copy of a complex object (CO) with write-permission and the complex object is changed on both sides. A complex object consists of objects that conceptually belong together in the sense of real-world modeling [4].

Complex objects  $CO_s$  in the server, contain information from different tables A, B, ..., Z which are known as child-object of aA, aB, ..., aZ, with aA being the root sub object. To explain the process of synchronization, let's define with:

$CO_1$  Copy of CO on client 1.

$CO_2$  Copy of CO on client 2.

$t_{d1}, t_{d2}$  are point in time when  $CO_1, CO_2$  are created/downloaded [4].

$t_{c1}, t_{c2}$  point in time when sub object aI of  $CO_1$  and sub object aJ of  $CO_2$  are changed  $\Rightarrow CO_1'$  and  $CO_2'$  with  $t_{c1} \neq 0$  and  $t_{c2} \neq 0$  (changes occurred on both sides).

$t_{s1}, t_{s2}$  - point in time when  $CO_1', CO_2'$  are synchronized.

$t_{cs}$  point in time when complex object is changed on server. Precondition:  $t_{d1} \leq t_{k1} \leq t_{s1}$  as well as  $t_{d2} \leq t_{k2} \leq t_{s2}$  and if  $t_{s1} \leq t_{s1}$ , client 1 synchronizes before client 2, respectively  $t_{cs} = t_{s1}$ , when client 1 synchronizes, its version is saved on the server. Conflict will be detected at time  $t_{s2}$  when trying to set  $t_{cs}$  again without having seen the former version. The detection of a conflict is possible using number ranges for unique identification. As every sub object, the root sub object of a complex object has a unique ID.  $CO_1$  and  $CO_2$  are copies of the same complex object if the ID of the root sub object is identical.

The allowed operations are Insert, Update and Delete. Delete is only allowed in a few tables which are checked out by the client. The conflict scenarios Update-Delete and

Delete-Delete are therefore not possible. Through the synchronization process the entries in the table are sent from client to server, executed on server side and stored in the server table. When another client downloads data it will receive the new entries from the table and can delete the records locally. Additionally, an Insert-Insert conflict is not possible either. Since every client and server has an own number range, the IDs will not be violated during synchronization. The only possible conflicting operation is Update-Update.

#### 4. Electronic Student Management System

With the process of synchronization of data it is possible to unify the data that are held in particle databases of special physical units. When one application modifies the data in one database, relevant changes pass to other databases. In our case, we have a central database and localized database in 17 faculties. During the transmission of data, from the central database to faculties, especially after the offline work of special faculties, we should take care how we treat data movements in order to avoid potential data conflicts.

ESMS is an UP system that is developed and implemented in all faculties of UP. System is Web based, with Web application and the database in the centre, and with the Web interface for the users in all faculties. The UP has built a network that connects all faculties. In UP we still have some problems with infrastructure reconstruction. The main problem remains with the electrical power interruption, since there are still difficulties to provide with a non-stop electrical power supply all the institutions and citizens in Kosovo. The influence of power supply is a major issue to the software systems. The electrical power interruption often happens during the work of the institutions so if the officials from the specific institution is using the system, the power interruption stops the process.

This problem is of direct consequence to the officials, since they need to provide services to the students who are expecting the availability of the information all the time. This leads to a diminished confidence of the students towards the faculty. Also the officials within the institutions decrease the confidence in UP systems. Also, since the officials in some cases are not very much familiar with the usage of different computerized systems, they tend hesitate working with new systems. The electrical power interruption has its own contribution that helps in the hesitance to adapt with new technologies. The Figure 2 shows the configuration of the servers and applications in the UP IT Centre and in one faculty.

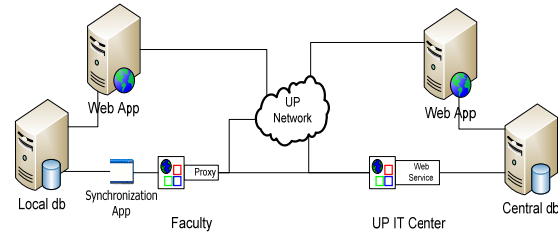


Fig. 2 Server configuration in UP IT Center and in one Faculty.

Taking into consideration these aspects and in order to succeed with this project, we had to find a solution that is independent from these problems. The solution was to introduce offline mode of operation. First thing to do in this case was to supply all faculties of UP with database and Web servers. Then, after the hardware infrastructure was completed, we had to install a copy of the application and the copy of the database in these servers. In this way, each of the faculties had the application at their disposal and could work with some of the features of the system. The problem lays still on the features of the system regarding the interconnection of other faculties and with the UP: How to synchronize the data from one side to another? We had to develop a special software component that does the data synchronization from one side to another.

##### 4.1 Technologies used to develop ESMS

ESMS is developed using modern technologies, which are numbered in Table 1.

Table 1: Technologies used to develop ESMS

No.	Description	Technology used
1	Server Operating	Microsoft Windows Server 2008™
2	Client Operating System	Windows, Linux and mobile
3	Development Platform	.NET Framework 3.5
4	Database	MS SQL Server 2008
5	Web Server	MS IIS 7
6	Browser	Internet Explorer, Mozilla Firefox, Opera, Safari, Opera
7	Programming languages	ASP.NET, C#, AJAX , HTML, CSS, JavaScript
8	Development Tools	Microsoft Visual Studio .NET 2008
9	Accessing Data	ADO.NET
10	Communication with other app.	Web Service, XML

## 5. Algorithm for data synchronization in ESMS

In UP there are still problems with infrastructure reconstruction. Main problems are: non-regular power supply, connection drops, server drops, which present a big problem in software applications. Such problems cause activity interruption at work and inability to do the service on time, which reflects the service to the students. Situations of this nature; cause skepticism to students, personnel and management of UP in use of IT services to the students. If there's connection, ESMS works parallel online and offline, which means the data are transferred in both local and central databases. If the connection is lost, ESMS works with local server, which means that new data are being saved in local server which are not in the central server and in this case the synchronization component should synchronize the data with the data center when connection is present. Web Service has the information on how data should be synchronized through the columns in the tables.

Considering those aspects and for the continuity of this project, a solution to minimize the problems should be found. For this purpose an offline mode is used, realized with the design of the algorithm used for data synchronization based on Web Services, as shown in Figure 3.

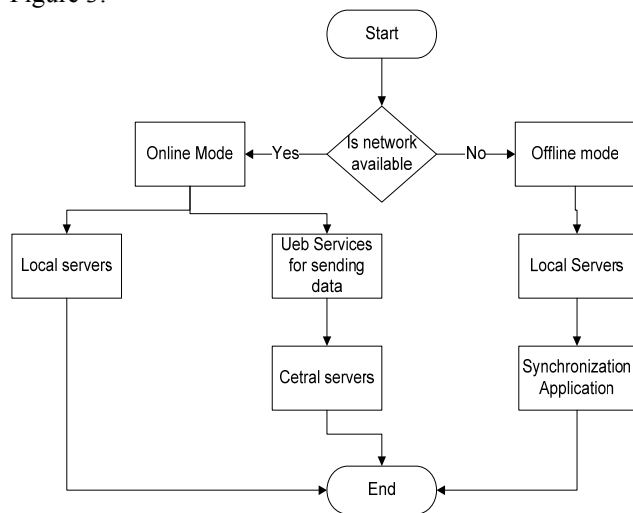


Fig. 3 Algorithm for data synchronization in SEMS .

## 6. Implementation of the proposed solution for ESMS in Offline mode

Implementation of Web applications in fact represents the online system, which will be set in a real Web server and

connected to the Internet. Some organizational units may not have access all the time to the Web applications due to their disconnection from the network. Those units that are not included in the online Web application will be enabled during the offline mode. The offline system could be a reduced version of the entire system or an online system with rights and obligations to users. However, the data stored in the system during offline operation, in case of reactivation of the network connection will enable the data synchronization with the online system, which will regret the role as primary system.

To provide both modes of operation, ESMS has built a system which enables the data synchronization between faculties that are working offline with the primary system, in the main data base, so that the data could still be synchronous and up to date with the work done in all the different units. For this purpose, it has been provided the data synchronization, for the units working in "Offline" mode, with the primary system. This synchronization is also based on techniques that make use of WSs, considered as a relatively new technology, and the network infrastructure that will support the implementation of such technology serving the UP and its data synchronization in both "Offline" and "Online" mode and vice versa. The real database will be the same one used from the "Online" mode of ESMS. In this case a WS will have one or more methods which will be provided only from the configuration of the web server at the UP. This is because the synchronization of the units that will be working "Offline" within the UP network will not have access to other networks outside the UP.

In order to make possible the connection between the different units throughout a WS a Proxy Client will be created, such that when a certain faculty will access the University network the data synchronization will be possible. The role of the proxy client would be to call the WS methods in order to transfer all the data that has been stored Offline and to receive updated data from the real database in the case when the unit has had no opportunity to communicate with central database when the interruption of the network had occurred. Besides the data transfer from the databases that work offline, units will also transfer files to the central server, from those units that have been working offline, when the network connection will back on. Files will be transferred as data and not as files, but the central server will return the file format. The data synchronization of the local and central server will be done in real time. Within a specified deadline a unit will be able to determine the Proxy activation from the Windows OS and the synchronization dynamics will depend from the needs that vary from hours, days, etc. In general, data and file synchronization

is possible, but the best techniques possible in order to get the best solution have been provided from a very close cooperation with the service provider at the UP.

### 6.1 Data Synchronization from Client

If the network is interrupted, we switch on the offline mode solution as it is presented in pseudo-code; the data are saved only on the local database server. In this case the synchronization components start and check if the network is present. When the network becomes present, this component calls WS that sends data that has to be synchronized with the central server [10]. Below is presented the pseudo-code used for data synchronization from the client with the respective comments.

```
private void Synchronization ()
{
    <WebServices> objWebService = new <WebService>();
    Synchronization Faculty >> UP_IT_Center
    for (int t = 0; t < numberOfTablesForSynchronization; t++)
    {
        Filling DataSet with faculty data for synchronization
    }
    Sendig of data to WS, that returns two DataSets:
    DataSet for confirmation that data are synchronized
    DataSet with data that has to be synchronized from server
    center in faculty
    <dsFromServerCenter[2]> = objWebService.Synchronization
    (<ID>,<Data>);
    Confirmation for data synchronization
    for (int t = 0; t < numberOfConfirmedTables; t++)
    {
        Confirmation that data into respective tables in faculty db
        are synchronized with the server in the centre
    }
    Synchronization UP_IT_Center>> Faculty
    Dat Synchronization that are sent from Server in the centre
    for (int t = 0; t < numberOfConfirmedTables; t++)
    {
        Registration or modification of send data
        Filling of DataSet for confirmation of synchronized data
    }
    Sending data for confirmation in server in the centre
    objWebService.Konfirmimi(ID, <Data for confirmation>);
    End of client synchronization
}
```

### 6.2 Data Synchronization from Server

In the next block, we have presented the pseudo-code for dat synchronization from the server in the center.

```
[WebMethod]
public DataSet[] Synchronization (string <ID>,
    DataSet <dsDataFromFaculty>)
{
    Synchronization Faculty >> UP_IT_Center
```

```
Data synchronization that are sent from Faculty
for (int t = 0; t < numberOfTablesForSynchronization; t++)
{
    Registration or modification of send data
    Filling of DataSet for confirmation of synchronized data
}
Synchronization UP_IT_Center>> Faculty
for (int t = 0; t < numberOfTablesForSynchronization; t++)
{
    Filling DataSet with data synchronization for server in the
    centre to server in faculty
}
<dsDataForFaculty[0]> = < Records to be synchronized >;
<dsDataForFaculty[1]> = < Data for confirmation >;
Returning of the data for synchronization and confirmation
in faculty
return < dsDataForFaculty >;
}
[WebMethod]
public void Confirmation(string <ID>, DataSet
<dsDataFromFaculty>)
{
    Confirmation of synchronized data in faculty
    for (int t = 0; t < numberOfConfirmedTables; t++)
    {
        Confirmation of data in respective tables in db in UP IT Centre
        that are synchronized with faculty
    }
}
```

## 7. Enrollment module and candidate acceptance in UP with ESMS

Enrollment module of candidates applying for registration in UP is implemented with purpose of avoiding manual tasks, which was practiced up to the year 2010. With this module time is reduced, cost and service quality is increased, correct and unified data are available divided for each faculty. Module for students acceptance within ESMS gives ability to see an overview of given results of candidates for the enrollment in UP, which includes: scores from preliminary (high school) studies, scores from national mature test and scores from entrance exam in faculty. Candidates can verify their personal data; receive exact information online for place, hall and time of entrance exam and final results for enrollment at all times. For the first time term students enrollment in UP in academic year 2011/12, all candidate have applied online using ESMS, from all around the world. The applicants' number has reached 17139 candidates, whilst 15734 applicants took the written test.

Statistics extracted from online applicants at all times from dbESMS database is done with the following query:

```
(SELECT COUNT (ApplicationID) FROM dbo.tblApplication
WHERE Fshije <> 1 AND ExamID > 49 AND FakultiyID =
NrFac) AS 'Name of Faculty' FROM tblApplication A.
```

Whilst for verified statistics from ESMS at all times from dbESMS database is done with the following query:

```
(SELECT COUNT (VerificationID) FROM dbo.tblVerification
WHERE Fshije <> 1 AND afatiID > 49 AND FakultiyID =
NrFac) AS 'Name of Faculty' FROM tblVerification V.
```

Figure 4 shows the diagram on which statistical data can be seen for numbers of applicant and those verified, from all faculties in UP, an ESMS's extraction.

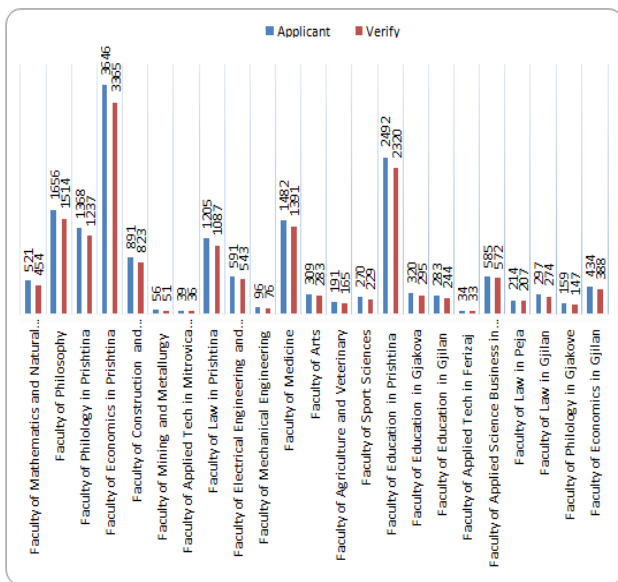


Fig. 4 Proposed beam former Online Applicants Statistics by faculty extracted with ESMS on the first acceptance term of students in UP for the academic year 2011/12.

## 8. Conclusions

In this paper we have studied and presented the importance of using data synchronization in an UP project called ESMS. In our study, we have investigated the WS oriented approach for data synchronization of shared data systems, enabling software applications to work in offline mode and thus increasing the confidence in work. We used synchronization to avoid problems that the UP of Prishtina have with the interruption of electricity and the network failures. By the use of synchronization, the officials in the institutions have no more problems with the network failures, and so the system functions without any interruption. Also the confidence of the administrative and academic staff in the UP increases since there is no

more waiting for the documents because the network fails. The synchronization created has shown positive results in the reliability of the users of the software applications. The synchronization that we have build for the solution of the problem on UP has found application in many electronic systems which are being developed within the e-Government in Kosovo [12]. Nowadays and in the future, system developers have to take into consideration the use of synchronization methods in order to efficiently implement their systems. In the future, application developers should consider using those methods for synchronization presented in this paper, to implement their systems with full efficiency.

## Acknowledgments

Authors are grateful for all participants that helped obtained data needed to make this paper. Special thanks go to Management of UP and to the team of IT Office in UP for their contributions in the implementation, testing and administration of ESMS.

## References

- [1] C. George, D. Jean, K. Tim, and B. Gordon, "Distributed System, Concepts and Design", Fifth Edition, Place: published by Addison Wesley, 2011.
- [2] J. N. Foster, and K. Grigoris, "Provenance and Data Synchronization", IEEE Data Engineering Bulletin, 2007, Vol. 30, pp. 13–21.
- [3] K. Tomas, and S. Jakub, "Lessons Learned on Enhancing Performance of Networking Applications by IP Tunneling through Active Networks", International Journal on Advances in Internet Technology, Vol.3, No. 3 & 4, 2010, pp. 223-233.
- [4] D. Dirk, and N. Christine, "A Context-Oriented Synchronization Approach", Software Competence Center Hagenberg, Softwarepark 214232 Hagenberg, Austria, PersDB, 2008, pp. 20-27.
- [5] J.B. Broka, "The Future of the Internet: Scenarios and Challenges in the Evolution Path as Seen by EIFFEL Think-Tank", International Journal on Advances in Internet Technology, Vol.3, No. 3 & 4, 2010, pp. 195-202.
- [6] N.S. Satish, and S. Vladimir, V. Vainikko, and J. Matthias, "Supporting Mobile Web Service Provisioning with Cloud Computing", International Journal on Advances in Internet Technology, Vol.3, No. 3 & 4, 2010, pp. 261-273.
- [7] A.M. Eyhab, and H. M. Qusay, "Investigating Web Services on the World Wide Web", Department of Computing and Information Science University of Guelph, Guelph, ON, N1G 2W1, Canada, 2008.
- [8] J.N. Foster, M.B. Greenwald, C. Kirkegaard, B.C Pierce, and A. Schmitt, "Exploiting Schemas in Data Synchronization", 2007.
- [9] D. Lieven, J. Bart, P.Frank, and J. Wouter, "Threat modelling for web services based web applications", DistriNet Research Group, Katholieke Universiteit Leuven, Celestijnenlaan 200A, 3001 Leuven, Belgium, 2005.



- [10] P. Wohed, W.M.P Aalst , M. Dumas, and A. Hofstede, "Analysis of Web services Composition languages: The case of BPEL4WS", Proc. of 22nd International Conference on Conceptual Modeling, Chicago, IL, 2003, pp. 200-215.
- [11] W.M.P. Aalst, and H.Verbeek, "Process Mining in Web Services: The WebSphere Case", IEEE Data Eng. Bull. 31(3), 2008, pp. 45-48.
- [12] "E-government strategy 2009-2015", <http://kk.rks-gov.net>.

**First Author: Isak Shabani**, Dipl. Ing. In Computers and Telecommunication – 2002, Mr. sc. in Computer Engineering – 2007; PhD Candidate, Computer Engineering; 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 Web Services and data synchronization.

**Second Author: Betim Cico**, Prof. Dr. Betim Cico was graduated as electronic engineer in Polytechnic University of Tirana (PUT) in 1970 with excellent results. After one year of working as electronic engineer in Shijak Radiostation, Albanian Radio-Television, he worked for 26 years as scientific researcher at the Institute of Nuclear Physics(INP), mainly in the field of microprocessors and computers, computer based on-line systems, Head of the Electronics Department, part time professor in PUT. In 1998 he moved as Head of Computer Engineering Section Electronic Department, Faculty of Electrical Engineering at PUT. During 1999-2011 member of the main governing body (Senate) at PUT. Participated in 50 International Trainings, Workshops, Conferences etc. Author of 40 Scientific Papers and 20 INP Technical Reports. Supervisor of three PhD Theses. During 1999-2001 member of the Project Group, MOES, for the implementation of the Education Management Information System (EMIS) Component under the Transition Education Reform Project In Albania. Designer of several LANs in Tirana.

**Third 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.

# Inversion of Web Service Invocation using Publish/Subscribe Push-Based Architecture

Thanisa Numnonda<sup>1</sup> and Rattakorn Poonsuph<sup>2</sup>

<sup>1</sup> School of Applied Statistics, National Institute of Development Administration,  
Bangkok, 10240, Thailand

<sup>2</sup> School of Applied Statistics, National Institute of Development Administration,  
Bangkok, 10240, Thailand

## Abstract

Among enterprise application integration solutions, Web services technologies are promising technologies to achieve the interoperability in heterogeneous environments. However, traditional Web service invocation may lead to unnecessary network traffic, long response time, and bottleneck problems at service providers. While a publish/subscribe model provides an advantage of prompt notification which can eliminate unnecessary network traffic, its achievement in interoperability is limited. By integrating Web services technologies with a publish/subscribe model, a pull-based architecture and a push-based architecture are mentioned in this paper. The pull-based architecture uses the integrated solution based on traditional Web service invocation, still the bottleneck problems at service providers are likely to occur. Therefore, we propose an alternative, the push-based architecture which presents an innovative approach of using inversion of Web service invocation. Instead of letting service clients invoke services at service providers as usual, the service clients simply wait for updated information from the service providers. Experimental results showed that the response time was significantly minimized and the bottleneck problems at service providers were eliminated in the push-based architecture. Thus, service providers can be very small and thin in ubiquitous computing such as sensor or mobile devices.

**Keywords:** *Traditional Web Service Invocation, Inversion of Web Service Invocation, Publish/Subscribe, Pull-Based Architecture, Push-Based Architecture.*

## 1. Introduction

Sharing information between applications among or within enterprises is a major business strategy. There are several solutions and concepts for sharing information to streamline the business workflow and to apply in enterprise application integration (EAI). In recent years, service-oriented architecture (SOA) concept is significantly getting industry attention by using Web service technologies which are promising technologies to achieve interoperability in heterogeneous environments.

However, SOA may also create mesh connections to multiple applications within an enterprise which is difficult to maintain. In addition, traditional Web service invocation (traditional-WSI) may lead to unnecessary network traffic and long response time when service clients are required to periodically poll for updated information. Moreover, bottleneck problems at service providers may occur when service providers confront with numerous active polling from service clients simultaneously. An overview of the polling architecture using traditional-WSI is shown in Fig. 1.

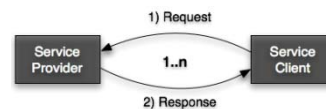


Fig. 1 An overview of the polling architecture using traditional-WSI.

Instead of using periodically polling mechanism, message-oriented middleware (MOM) and a publish/subscribe (pub/sub) model can be used for prompt notification which can eliminate unnecessary network traffic. After service providers and service clients register as publishers and subscribers respectively, all registered service clients can be notified of updated information when available. By integrating Web services technologies based on traditional-WSI with a pub/sub model, a pull-based architecture is used in the way that after service clients receive notification message, they have to send requests to service providers in order to get updated information. Therefore, when there are numerous requests simultaneously, the bottleneck problems at service providers are still likely to occur. An overview of the pull-based architecture using traditional-WSI with a pub/sub model is shown in Fig. 2.

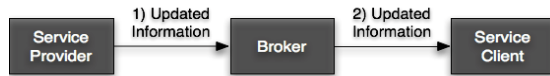


Fig. 2 An overview of the pull-based architecture using traditional-WSI with a pub/sub model.

Finally, we propose an alternative, a push-based architecture which is a complete solution for resolving shortcomings of EAI. The push-based architecture presents an innovative approach of using inversion of Web service invocation (inversion-WSI) with a pub/sub model. After service providers push Web service messages to the broker, those messages are propagated to all registered service clients. Therefore, instead of letting service clients invoke services at service providers as usual, service clients simply wait for updated information from service providers. Apparently, the response time can be reduced since service clients are able to receive updated information once available without sending any requests. The bottleneck problems at service providers are eliminated as the broker, designed to acquire high performance, is responsible for handling all services instead. Thus, service providers can be very small and thin in ubiquitous computing such as sensor or mobile devices. An overview of the push-based architecture using inversion-WSI with a pub/sub model is shown in Fig. 3.

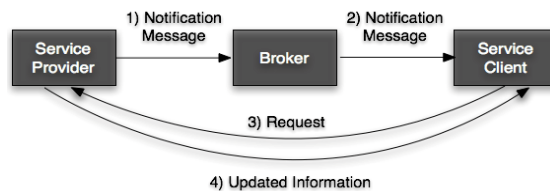


Fig. 3 An overview of the push-based architecture using inversion-WSI with a pub/sub model.

The rest of this paper is organized as follows. Section 2 provides some backgrounds on a pub/sub model and Web services technologies. Section 3 describes related works. Section 4 explains conceptual models of inversion-WSI, pull-based architecture, and push-based architecture. Section 5 clarifies research methodology in two approaches: mathematical models for total response time of pull-based and push-based architectures including comparison between them and implementation of the push-based architecture in details. Section 6 shows performance comparison results between pull-based and push-based architectures. Finally, section 7 discusses conclusion and future work.

## 2. Background

To share information or process among or within enterprises, EAI of heterogeneous systems is inevitable. Heterogeneous systems are normally developed by using several computer programming languages, different technologies, and deployed on various platforms. Therefore, integrating them is non-trivial. In the mid-1990s, evolution of EAI was started as enterprises tried to integrate the systems by using point-to-point connections between their applications [1]. It was successful in that era since there were only limited applications. However, the complexity of linkages between applications and difficulty of maintenance integration portions turned into problems when there were many applications. Additionally, data transformation and code conversion increased difficulty to implement. Therefore, several systematic approaches have been introduced to improve efficiency with minimal maintenance.

In the late 1990s, MOM became a very popular methodology used in EAI. The middleware concept is to allow applications to pass messages to others with single connection to MOM and more maintainable. MOM supports two types of communication: queue and topic. The queue in MOM can send a message to one consumer at a time whereas the topic in MOM with a pub/sub model is a better model and can send a message to multiple consumers concurrently [2]. This model usually consists of three basic elements: publisher, subscriber, and broker. The publisher is any application that wants to produce a message. The subscriber is any application registered to receive a copy of the message. The broker is the intermediary between publishers and subscribers. An application can be both a publisher and a subscriber at the same time. In EAI life cycle, number of publishers and subscribers can grow and shrink over time. Updated information can be either pulled by subscribers or pushed by publishers. Publishers can multicast a message of a topic to subscribers who subscribed on the topic via a broker [3]. Although publishers and subscribers are loosely coupled and transparent to each other, they are required to operate on the middleware infrastructure.

Web services technologies are promising technologies to achieve the interoperability in heterogeneous environments. An application often communicates with other applications using XML to encapsulate data and context. Using XML makes Web services platform, language, and vendor independent. As a result, Web services are ideal to be candidates for EAI solutions. Two main cores of first-generation Web services standard are SOAP (originally defined as Simple Object Access

Protocol) [4], a simple XML-based protocol and WSDL (Web Service Description Language) [5], an XML-based language to describe Web services. For traditional-WSI, any application wants to be a service provider must provide WSDL as a Web service interface. Service clients can then invoke services through stubs generated from the WSDL. An overview of traditional-WSI is shown in Fig. 4

Second-generation of Web services can form complex Web service applications. WS-\* deals with aspects such as security, transactions, messaging, and notification [6].

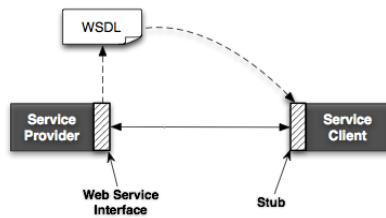


Fig. 4 An overview of traditional-WSI.

Enterprise service bus (ESB) [7], another approach to EAI, allows applications to communicate via a bus. The bus acts as a broker which basically supports multiple protocols such as being in the form of a pub/sub broker between message-based services. ESB is considered the next generation of MOM and it also extends functionality of MOM. However ESB normally requires extra level of translation which can decrease performance. As a result, we decide to use the integration of Web services technologies with a pub/sub model in this research due to its interoperability and performance for EAI.

### 3. Related Work

Some existing specifications and works already used the integration of Web services technologies as traditional-WSI with event notifications. Two competing specifications, Web Services Notification (WS-Notification) [8] and Web Services Eventing (WS-Eventing) [9], are crucial for asynchronous Web service-based event notifications. WS-Messenger supports both WS-Notification and WS-Eventing along with mediation between them. We briefly summarize WS-Notification, WS-Eventing, and WS-Messenger including some other related works in this section.

#### 3.1 WS-Notification

WS-Notification endorsed by OASIS, is to standardize the message protocols for topic-based or content-based pub/sub mechanisms based on Web services. There is a family of related three specifications: WS-

BaseNotification (mechanisms for basic notification), WS-BrokeredNotification (intermediary brokering capability), and WS-Topics (means to categorize notifications). WS-Notification implementation and extension are described in [10-12]. WS-BrokeredNotification is very similar to the pull-based architecture of this research.

#### 3.2 WS-Eventing

WS-Eventing is a new version and much simpler than WS-Notification. WS-Eventing basically relies upon WS-Addressing [13] for endpoint addresses. However, it only defines key pub/sub related functions such as subscribe, unsubscribe, and renew. Y.Huang et al. [14] compares WS-Notification and WS-Eventing in almost all aspects such as the delivery mode, message structure, and filter.

#### 3.3 WS-Messenger

WS-Messenger is a project from Indiana University [15]. It aims to support both WS-Notification and WS-Eventing specifications and conveys mediation between them by using Normalization-Processing-Customization (NPC) model. WS-Messenger can reduce some overheads in SOAP message processing since it processes SOAP messages directly at the XML message level without creating data binding between XML elements and Java objects. R. Jayasinghe et al. [16] presents few approaches motivated by WS-Messenger to improve message delivery of pub/sub system at the broker.

#### 3.4 Other Work

X. Feng et al. [17] used message-driven pub/sub system to help servers push recommended Web Services to customers based on subscribed conditions. The architecture for push-based Web service wrappers is focused by L. Brenna and D. Johansen [18], but it still uses the wrapper to regularly pull Web services. Therefore, some pull requests may return unchanged data which cause unnecessary network traffic and run down server resources. To the best of our knowledge, however, all of the related works we mentioned do not use the concept of inversion-WSI. Thus, they cannot eliminate the bottleneck problems at service providers whereas our push-based architecture can.

## 4. Conceptual Models

Since we use the concept of inversion-WSI in our push-based architecture, an overview of inversion-WSI is demonstrated in this section. To understand the response time comparison in the next section, sequence diagrams of

pull-based and push-based architectures are also provided here.

#### 4.1 Inversion of Web Service Invocation (Inversion-WSI)

Instead of letting service clients invoke services at service providers as usual, service clients simply wait for updated information from service providers. This is called inversion-WSI which is an opposite of traditional-WSI. The broker is responsible for defining the canonical message comprising of a name and WSDL of a topic. A service provider must generate a stub from the WSDL so that it can invoke a service at the broker through the stub. Meanwhile, service clients must also provide Web service interfaces of the same WSDL for the broker to invoke services. Therefore, to make inversion-WSI possible, all service providers (as publishers) and service clients (as subscribers) of the same topic must use the same canonical message. An overview of inversion-WSI is shown in Fig. 5.

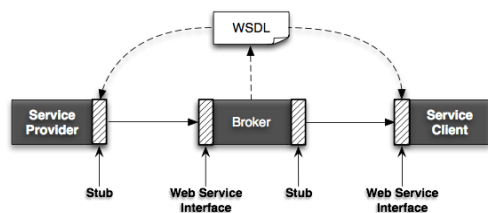


Fig. 5 An overview of inversion-WSI.

#### 4.2 Pull-based Architecture

This architecture is based on traditional-WSI. When a predefined event occurs at a publisher, the publisher sends a notification message to a broker. The broker will then propagate that notification message to all registered subscribers. After that, subscribers have to send requests to the publisher in order to get updated information. Finally, acknowledgements must be sent from subscribers to the broker so that the broker can keep track of successful or failed transmissions. A sequence diagram of the pull-based architecture is shown in Fig. 6.

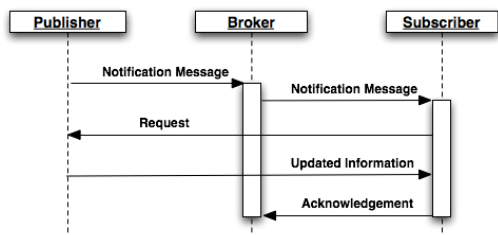


Fig. 6 A sequence diagram for the pull-based architecture.

However, this architecture has two shortcomings. First, the workload on publishers can be very high when they face numerous requests from subscribers simultaneously. Second, the response time is likely to be time-consuming since this architecture requires at least four one-way communications for a subscriber to get updated information.

#### 4.3 Push-based Architecture

This architecture is based on inversion-WSI. The transfer of updated information is triggered by a predefined event at a publisher. The publisher first pushes the updated information to a broker, then the broker multicasts that information to all corresponding subscribers. As a result, subscriber can receive updated information without sending any requests. Acknowledgements of subscribers must be sent to the broker in order to keep track of successful or failed transmissions. Fig. 7 shows a sequence diagram to describe this architecture.

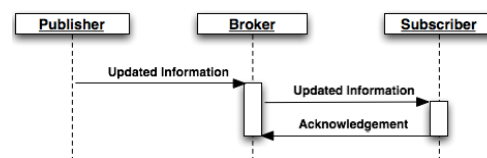


Fig. 7 A sequence diagram for the push-based architecture.

The push-based architecture is good for wide area distributed systems since publishers have no necessity to process numerous requests from subscribers. For this reason, publishers can be very small and thin. Besides, the response time for a subscriber to receive updated information is minimized to merely two one-way communications. We can briefly compare performance between three architectures shown in the Table 1.

Table 1: Performance comparison between three architectures

	<i>Polling Architecture (traditional-WSI)</i>	<i>Pull-based Architecture (traditional-WSI + pub/sub)</i>	<i>Push-based Architecture (inversion-WSI + pub/sub)</i>
Unnecessary Network Traffic	occur	is eliminated	is eliminated
Response Time	is long	is shorter	is shortest
Bottleneck Problems	occur	still occur	are eliminated

## 5. Research Methodology

To clarify our research methodology, mathematical models for total response time of pull-based and push-based architectures are provided and compared. Implementation details in two phases of the push-based architecture; pre-installation phase and runtime phase are also described in this section.

### 5.1 Mathematical Models

There are five steps of the pull-based architecture and only three steps of the push-based architecture to calculate the total response time. We approximate that processing time at a publisher and at a broker of sending a notification message are the same and equal to  $p_n$ . All important symbols and their meanings are listed in Table 2. Mathematical models of the pull-based and push-based architectures are shown in Fig. 8 and Fig. 9 respectively.

Table 2: Symbols and meanings

Symbol	Meaning
$p_n$	Processing time of sending a notification message
$p_{pi}$	Processing time at a publisher of sending updated information
$p_{bi}$	Processing time at a broker of sending updated information
$t_n$	Sending time of a notification message
$t_r$	Sending time of a request message
$t_i$	Sending time of updated information
$t_a$	Sending time of an acknowledgement message
$p_s$	Processing time at a subscriber

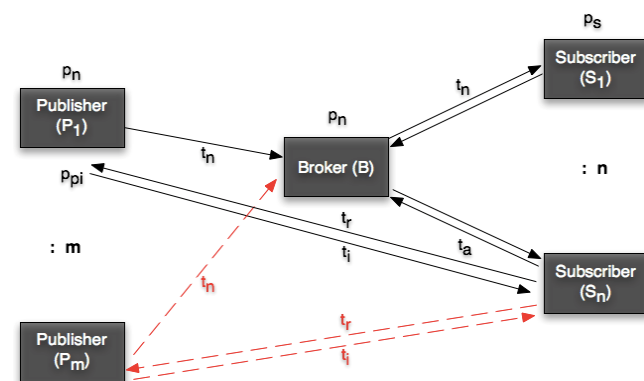


Fig. 8 A mathematical model of pull-based architecture.

**Step 1:** When a predefined event occurs at a publisher, the publisher sends a notification message to a broker. The summation of processing time at the publisher and sending

time of a notification message from the publisher to the broker (P to B) is defined as

$$p_n + t_n$$

**Step 2:** After the broker receives the notification message from the publisher, it forwards that message to all registered subscribers. The summation of processing time at the broker and sending time of a notification message from the broker to all registered subscribers (B to  $S_1... S_n$ ) is defined as

$$(n * p_n) + (n * t_n)$$

**Step 3:** In order to get updated information, a subscriber has to send a request message to the publisher. In this research, we assume that all registered subscribers send requests to the publisher. Therefore, the summation of processing time at the subscriber and sending time of a request message from all registered subscribers to the publisher ( $S_1... S_n$  to P) is defined as

$$(n * p_s) + (n * t_r)$$

**Step 4:** After the publisher gets a request, it will process the request and send updated information as a response. Sending time of updated information depends on size of updated information ( $s_i$ ), therefore  $t_i$  is equal to the time constant ( $\tau$ ) multiplied by  $s_i$ . The summation of processing time at a publisher and sending time of updated information from the publisher to all registered subscribers (P to  $S_1... S_n$ ) is defined as

$$(n * p_{pi}) + (n * t_i) \quad \text{where } t_i = \tau * s_i \quad (1)$$

**Step 5:** After receiving updated information from the publisher, a subscriber must return an acknowledgement message to the broker. The summation of processing time at the subscriber and sending time of an acknowledgement message from all registered subscribers to the broker ( $S_1... S_n$  to B) is defined as

$$n * t_a$$

By summarizing all the above five steps, the formula of the total response time for  $m$  publishers of the pull-based architecture ( $t_{pull}$ ) is defined as

$$t_{pull} = \sum_m (\text{step 1} + \text{step 2} + \text{step 3} + \text{step 4} + \text{step 5})$$

$$t_{pull} = m * ((p_n + t_n) + ((n * p_n) + (n * t_n)) + ((n * p_s) + (n * t_r)) + ((n * p_{pi}) + (n * t_i)) + (n * t_a))$$

$$t_{pull} = m * ((n+1) * (p_n + t_n)) + (n * p_s + t_r + p_{pi} + t_i + t_a)) \quad (2)$$

To simplify the formula (2) when n is large ( $n + 1 \sim n$ ), the summation of the step 1 (from P to B) can be ignored. Therefore, the formula (2) can be rewritten into the formula (3) as

$$t_{pull} \sim m * n * (p_n + t_n + p_s + t_r + p_{pi} + t_i + t_a) \quad (3)$$

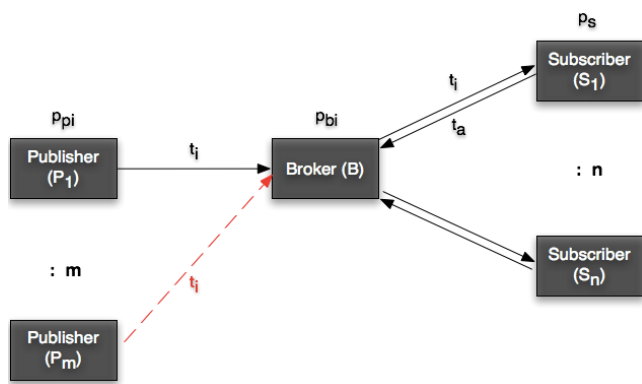


Fig. 9 A mathematical model of push-based architecture.

**Step 1:** When a predefined event occurs at a publisher, the publisher will send updated information to a broker. The summation of processing time at the publisher and sending time of updated information from the publisher to the broker (P to B) is defined as

$$p_{pi} + t_i$$

**Step 2:** After the broker receives updated information from the publisher, the broker will retrieve all endpoint addresses of registered subscribers and forward that updated information to them. Sending time of updated information depends on size of updated information ( $s_i$ ), therefore  $t_i$  is equal to the time constant ( $\tau$ ) multiplied by  $s_i$ . The summation of processing time at the broker and sending time of updated information from the broker to all registered subscribers (B to  $S_1 \dots S_n$ ) is defined as

$$(n * p_{bi}) + (n * t_i) \quad \text{where } t_i = \tau * s_i \quad (4)$$

**Step 3:** All registered subscribers can receive updated information from the publisher via the broker without sending any request. After receiving updated information, a subscriber should return an acknowledgement to the broker. The summation of processing time at the subscriber and sending time of an acknowledgement message from all registered subscribers to the broker ( $S_1 \dots$

$S_n$  to B) is defined as

$$(n * p_s) + (n * t_a)$$

By summarizing all the above three steps, the formula of the total response time for m publishers of the push-based architecture ( $t_{push}$ ) is defined as

$$t_{push} = \sum_m (\text{step 1} + \text{step 2} + \text{step 3})$$

$$t_{push} = m * ((p_{pi} + t_i) + ((n * p_{bi}) + (n * t_i)) + ((n * p_s) + (n * t_a))) \quad (5)$$

To simplify the formula (5) when n is large ( $n + 1 \sim n$ ), the summation of the step 1 (from P to B) can be ignored. Therefore, the formula (5) can be rewritten into the formula (6) as

$$t_{push} \sim m * (p_{pi} + (n * (p_{bi} + t_i + p_s + t_a))) \quad (6)$$

Comparing the difference total response time between  $t_{pull}$  from (3) and  $t_{push}$  from (6)

$$t_{pull} \sim m * n * (p_n + t_n + p_s + t_r + p_{pi} + t_i + t_a)$$

$$t_{push} \sim m * (p_{pi} + (n * (p_{bi} + t_i + p_s + t_a)))$$

$$t_{pull} - t_{push} \sim m * ((n * (p_{pi} - p_{bi})) + (n * (p_n + t_n + t_r + t_i))) \quad (7)$$

Let  $\delta$  be the difference processing time between at a publisher and at a broker of sending updated information,  $\delta = p_{pi} - p_{bi}$ , the formula (7) can be rewritten into the formula (8) as

$$t_{pull} - t_{push} \sim m * n * (\delta + p_n + t_n + t_r) \quad (8)$$

If the publisher and broker have the same specifications ( $\delta=0$ ), the difference total response time between the pull-based and push-based architectures will depend mainly on ( $p_n + t_n + t_r$ ) as shown in the formula (9). However, the processing time at the broker is normally much less than that of the publisher and if  $\delta$  is much greater than ( $p_n + t_n + t_r$ ), the difference time between the pull-based and push-based architectures is mainly depended on  $\delta$  as shown in the formula (10).

$$t_{pull} - t_{push} \sim m * n * (p_n + t_n + t_r) \quad \text{when } \delta=0 \quad (9)$$

$$t_{pull} - t_{push} \sim m * n * (\delta) \quad \text{when } \delta \gg (p_n + t_n + t_r) \quad (10)$$

## 5.2 Implementation of the Push-based Architecture

Implementation processes of the push-based architecture are set up and carried out based on the following conditions:

- There is only one operation in unique WSDL per topic.
- There can be multiple publishers per topic and a publisher can also be a subscriber of the same topic. It means that one or more publishers can simultaneously publish similar messages to subscribers of the same topic.
- When a publisher, at the same time acts as a subscriber of the topic, sends a message to the broker, the broker will handle this situation by not sending that message back to the publisher.

There are two phases of the push-based architecture needed to be explained: pre-installation phase and runtime phase. For the pre-installation phase, to be able to make inversion-WSI possible, any application interested to be a source of updated information must register as a publisher of a topic at a broker. After that, it can obtain WSDL of that topic from the broker. Any application interested to receive the updated information must get a related file to the WSDL of the topic from the broker and can implement in its desired way. Finally, interested subscribers must register and provide their endpoint addresses to the broker. An overview of the pre-installation phase of push-based architecture is shown in Fig. 10.

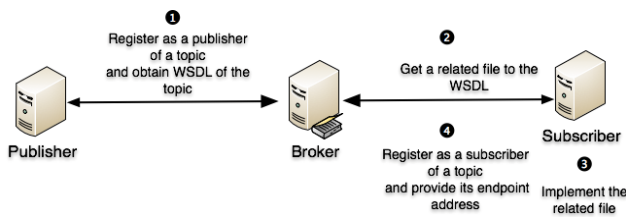


Fig. 10 An overview of the pre-installation phase of push-based architecture.

For the runtime phase, when a predefined event occurs at a publisher, the publisher invokes Web service at a broker and the broker then invokes Web services of all registered subscribers. Each subscriber must return an acknowledgement back to the broker so that the broker can keep track of successful or failed transmissions. An overview of the runtime phase of push-based architecture is shown in Fig. 11.

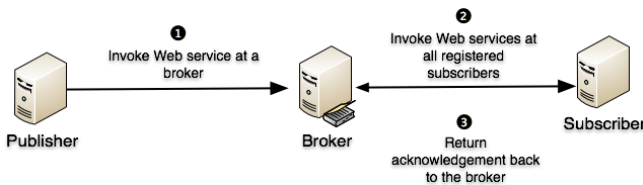


Fig. 11 An overview of the runtime phase of push-based architecture.

## 6. Experimental Results

In this section, experimental results are presented for performance comparison between pull-based and push-based architectures. The simulation of both architectures was set up within the same running environment such as the same processor speed and the same network bandwidth. We used twelve identically configured machines: Pentium 2.4 GHz and 2GB of RAM on Window 7 for both architectures. Ten machines were used for subscribers with up to 10 simulated subscribers on each machine. Each subscriber on the same machine was operated on a separated but identical server. A broker and a publisher each acquired own machine.

We started to measure the total response time after the broker received a notification message or updated information from the publisher. The summation time from the publisher to the broker was ignored as explained from the formula (2) to (3) of the pull-based architecture and from the formula (5) to (6) of the push-based architecture. The total response time would end after the broker received acknowledgements from all subscribers. We experimented in 20 times of each total response time and calculated the average of them.

Three scenarios were experimented to find the average total response time with following factors:

1. The number of subscribers was increased from 10 to 100 with a step of 10 by fixing the size of updated information to 1 Kbyte and using only 1 publisher.
2. The size of updated information was increased from 4Kbyte to 40Kbytes with a step of 4 by fixing the number of subscribers to 40 and using only 1 publisher.
3. The number of publishers was increased from 1 to 10 with a step of 1 by fixing the number of subscribers to 20 and the size of updated information to 1 Kbyte. Some publishers may also be subscribers of the same topic.

Experimental results of the first scenario are shown in Fig. 12. When the number of subscribers was increased from 10 to 20 and continuously into 100, we found that average total response time of the pull-based architecture was rising higher than that of the push-based architecture. Since the same specification of machine was used for the broker and the publisher ( $\delta=0$ ) and the number of publisher was fixed to 1 ( $m=1$ ),  $(p_n + t_n + t_r)$  in the formula (9) could be approximated to 10 ms. Therefore, the number of subscribers ( $n$ ) is influential to difference total response time between the pull-based architecture and the push-based architecture by around  $n * 10$  ms.



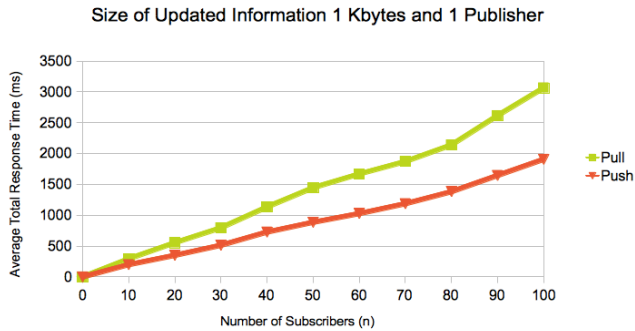


Fig. 12 Average total response time when the number of subscriber was increased.

Experimental results of the second scenario are shown in Fig. 13. When the size of updated information was increased from 4Kbytes to 8Kbytes and continuously into 40Kbytes, we found that difference average total response time between the pull-based architecture and the push-based architecture remained nearly the same. From the formula (3) and (6), both architectures need to send updated information ( $t_i = \tau * s_i$  in the formula (1) and (4)), thus the difference time does not rely on the size of updated information. In this scenario, we set up 4 subscribers per machine for 10 machines to be the total of 40 subscribers.

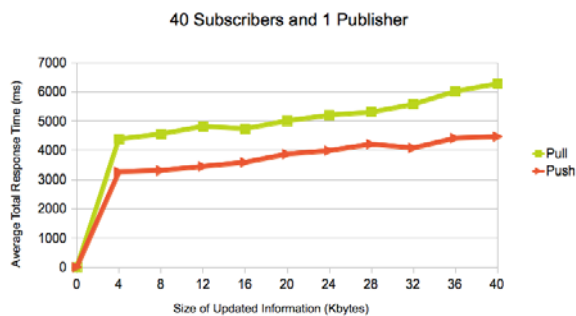


Fig. 13 Average total response time when the size of updated information was increased.

To be able to measure the average total response time from many publishers, lots of notification messages or updated information were sent out from publishers in the pull-based and push-based architecture respectively. Experimental results of the last scenario are shown in Fig. 14. When the number of publisher was increased up to 5, the average total response time of the pull-based architecture was higher than that of the push-based architecture. However, when the number of publisher went

beyond 5, bottleneck problems occurred in the pull-based architecture which caused the total response time could not be determined. The main reason of the bottleneck problems came from that some publishers who at the same time acted as subscribers were not able to handle concurrent Web services invocation properly. In this scenario, we set up 2 subscribers per machine for 10 machines to be the total of 20 subscribers.

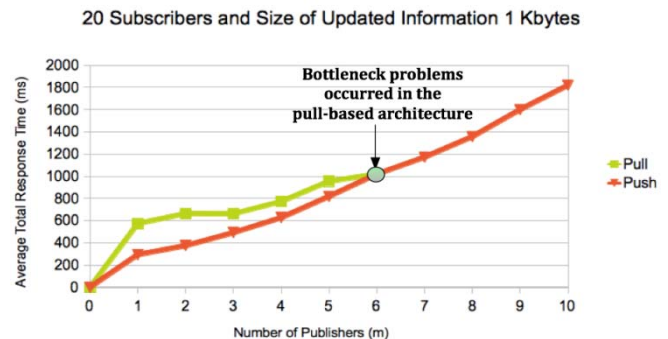


Fig. 14 Average total response time when the number of publisher was increased.

## 7. Conclusion and Future Work

The results of this research show that the push-based architecture surpasses the pull-based architecture by integrating Web services technologies with a pub/sub model. Using inversion-WSI instead of traditional-WSI, the broker is a core component of the push-based architecture. It acts as a middleware for receiving and sending updated information, as well it may perform several functions such as data transformation, code conversion, and conditional routing. Therefore, the broker should be designed to be able to handle heavy workloads whereas service providers can be very small and thin.

Since the push-based architecture can significantly minimize overall response time and workload on service providers, it is potentially applicable for some machine-to-machine (M2M) applications that need to speedily distribute updated information in urgent situation such as Tsunami alert system. The push-based architecture can efficiently support tracking updated information for many purposes as well.

Although this paper does not mention about the security, quality of service (QoS), and transaction, the concept of Web Services Atomic Transaction (WS-Atomic Transaction) [19] can be applied to enhance the reliability which is considered to be our future work. For further work, service clients may be able to choose which data

they want to receive via the pull-based architecture or the push-based architecture. The factor to choose between architectures may depend on a category or size of updated information. This may be called hybrid Web service invocation (hybrid-WSI).

### Acknowledgments

This research was partially funded by the National Institute of Development Administration (NIDA). We also would like to thank the School of Applied Statistics of NIDA for providing us supports on this research work.

### References

- [1] J. Lee, K. Siau, and S. Hong, "Enterprise integration with ERP and EAI", in Communications of the ACM, 2003, 46 (2), pp. 54–60.
- [2] P. T. Eugster, P. A. Felber, R. Guerraoui, and A.-M. Kermarrec, "The Many Faces of Pub/sub", in ACM Computing Surveys, 2003, 35 (2), pp. 114-131.
- [3] I. Gorton, Essential Software Architecture, Springer, 2006.
- [4] W3C, SOAP Version 1.2 Part 1, Available: <http://www.w3.org/TR/soap12-part1/>
- [5] W3C, Web Services Description Language (WSDL) Version 2.0 Part 1, Available: <http://www.w3.org/TR/wsdl20/>
- [6] Erl, Thomas, SOA Principles of Service Design, Service Oriented Computing Series, Prentice Hall, 2007.
- [7] J. Wu and X. Tao, "Research of Enterprise Application Integration Based-on ESB", in 2nd International Conference on Advanced Computer Control (ICACC), 2010.
- [8] OASIS, Web Services Notification (v 1.3), Available: <http://docs.oasis-open.org/wsn/>
- [9] W3C, Web Services Eventing, Available: <http://www.w3.org/Submission/WS-Eventing/>
- [10] M. Humphrey, G. Wasson, K. Jackson, J. Boverhof, M. Rodriguez, Joe Bester, J. Gawor, S. Lang, I. Foster, S. Meder, S. Pickles, and M. McKeown, "State and Events for Web Services: A Comparison of Five WS-Resource Framework and WS-Notification Implementations", in 4th IEEE International Symposium on High Performance Distributed Computing (HPDC-14), 2005.
- [11] A. Quiroz and M. Parashar, "Design and Implementation of a Distributed Content-based Notification Broker for WS-Notification", in Grid Computing Conference, 2006.
- [12] S. D. Labey and E. Steegmans, "Extending WS-Notification with an Expressive Event Notification Broker", in 2008 IEEE International Conference on Web Services, 2008.
- [13] W3C, Web Services Addressing 1.0 – Core, Available: <http://www.w3.org/TR/ws-addr-core/>
- [14] Y. Huang and D. Gannon, "A Comparative Study of Web Services-based Event Notification Specifications", in Proceedings of the 2006 International Conference on Parallel Processing Workshops (ICPPW'06), 2006.
- [15] Y. Huang, A. Slominski, C. Herath, and D. Gannon, "WS-Messenger: A Web Services-based Messaging System for Service-Oriented Grid Computing", in Proceedings of the Sixth IEEE International Symposium on Cluster Computing and the Grid (CCGRID'06), 2006.
- [16] R. Jayasinghe, D. Gamage, and S. Perera, "Towards Improved Data Dissemination of Publish-Subscribe Systems", in 2010 IEEE International Conference on Web Services, 2010.
- [17] X. Feng, F. Xue, and T. Zhang, "Research on data exchange push technology based on message-driven", in 2009 International Joint Conference on Artificial Intelligence", 2009.
- [18] L. Brenna and D. Johansen, "Configuring Push-Based Web Services", in Proceedings of the International Conference on Next Generation Web Services Practices (NWeSP'05), 2005.
- [19] OASIS, WS-AtomicTransaction (v 1.2), Available: <http://docs.oasis-open.org/ws-tx/wstx-wsat-1.2-spec.html>

**Thanisa Numnonda** is a Ph.D. candidate at the School of Applied Statistics, National Institute of Development Administration, Bangkok, Thailand. She received a Master's Degree in Computer Engineering from the Department of Electrical and Computer Engineering, University of Southern California, USA. Her research interests are in Web Services and Service-Oriented Architecture.

**Rattakorn Poonsuph** is a lecturer at the School of Applied Statistics, National Institute of Development Administration, Bangkok, Thailand. He received a Ph.D. degree in Computer Science from University of Massachusetts Lowell, USA. His fields of research include Software Engineering and Software Architecture. He has published several papers in various international conferences.

# Design of a Conceptual Reference Framework for Reusable Software Components based on Context Level

V. Subedha<sup>1</sup>, Dr. S. Sridhar<sup>2</sup>

<sup>1</sup> Research Scholar, Department of CSE, Sathyabama University  
Chennai, Tamilnadu, India

<sup>2</sup> Research Supervisor, Department of CSE, Sathyabama University  
Chennai, Tamilnadu, India

## Abstract

Reusable software components need to be developed in a generic fashion that allows their reusability in context level. Components identification based on quality metrics for reusability and indexing had been the desired technique in the field of reusable software components. However, the methodologies utilized for the identification of reusable components are not able to handle the reusability of faulty behavior component. In this paper we propose a conceptual reference framework for reusable software components which is available in reusable component repositories and also based on the faulty functional behavior of the components in the environment. Also we propose a Component Extraction scheme named as Minimum Extraction Time First (METF) based on extraction time of the component. The component for reuse is qualified based on the functional coverage report, software reuse metrics and minimum extraction time from the collection of components identified. Reuse-Utility-Percent and Reuse-Frequency metrics were used to assess the reusability in the environment. So, the proposed framework can be used to achieve high potential and high quality reuse.

**Keywords:** *Software reuse, Reusable Software Components, Components Identification, Component Extraction, Component Qualification, Reusability metrics.*

## 1. Introduction

Effective reuse of requirements, architecture, design, process, technology, knowledge and components from previous software developments can increase the productivity & quality in software environment [4]. Software reuse catalyzes improvements in productivity by avoiding redevelopment and improvements in quality by incorporating components whose reliability has already been established [7]. In fact, Software production using the reusable components will probably be very crucial to the higher level of software industry maturity [9]

So, in recent years, there has been an increasing awareness of the reusability. As a consequence there are lots of

research studies which focus on the software components extracted from the existing sources [1]. The most extensive effort to date has been the focus on software reuse is to examine the issues ranging from methods and techniques and also to improve productivity and economic impact [2].

Reusability can be improved not only by replicating the software components but also continue to reuse the components with faulty functional behavior. Coverage driven functional verification plays a more and more important role in reusability of components with faulty behavior. Based on the coverage report generated by the Coverage driven test-cases the components are identified for reusability.

Component-based reuse is widely accepted as an important reuse strategy and component-based reuse programs heavily depend on software reuse repositories for achieving success [12].

Reuse Frequency metric and Reuse-Utility-Percent metric values are used as the assessment attributes for reusability of the software component in Context level. In this paper, we outline a way to reuse the software components corresponding to the context-dependent reusability.

The rest of this paper is structured as follows. Section 2 deals with some of the related research works. Section 3 describes our design of a conceptual reference framework for Reusable Software Components and the proposed algorithm for the framework. Section 4 deals with metrics to assess reusability and their evaluation. Section 5 discusses about the case study and analysis results. Finally, Section 6 concludes.

## 2. Background and Related Work

In the literature, the area of studies addressed by this paper is called Reusable Software Components. This term means the process of effectively reusing the components from existing environment on context level.

Reusability is an important goal in well engineered systems. Two main approaches have been developed to identify the components for reusability in existing software components. The first one facilitates the component identification based on indexing and the second approach based on metrics and models.

P. Vitharana et al.[11] describes a component retrieval mechanism that the reuser formulates a request based on the keyword. To be retrieved, the information about the components must be encoded. This process of identification is also called as indexing which may be manual or automatic but result in loss of information.

Richard W. Selby [7] proposed a Reuse based software development model in order to achieve an average reuse of 32% per project. They identify two categories of factors that characterize successful reuse-based software development of large-scale systems: module design factors and module implementation factors. Also they evaluate the fault rates of the reused, modified, and newly developed modules.

Matteo Gaeta et al. [1] proposed an approach to extract relevant ontology concepts and their relationships from a knowledge base of heterogeneous text documents. The proposed model is a complete methodology for automatic knowledge extraction for reuse.

Parvinder Singh and Sandhu and Hardeep Singh, [5] [6] have used metric based approach for identifying a software module and the reusability was obtained with the help of Fuzzy Logic and Neuro-Fuzzy. This research shows how metrics can be used to identify the quality of a software component.

Keiji Hokamura et al [3] proposed an approach of defining reusable components for multiple Web applications using a domain-specific aspect-oriented (AO) mechanism based on an abstraction model common to all Web applications. The domain-specific AO mechanism based on the server-client model of Web is useful to describe reusable components which implement functionalities affecting user page accesses. B. Morel et.al [10] proposed adapting software components at the architecture level.

### 3. Conceptual Reference Framework for Reusable Software Components

Our reference framework for reusable software components is to identify, extract, qualify and integrate reusable software components based on the functional behavior. This approach follows the well planned, efficient by cost and quality product. Fig. 1 shows the conceptual reference framework for reusing the software components.

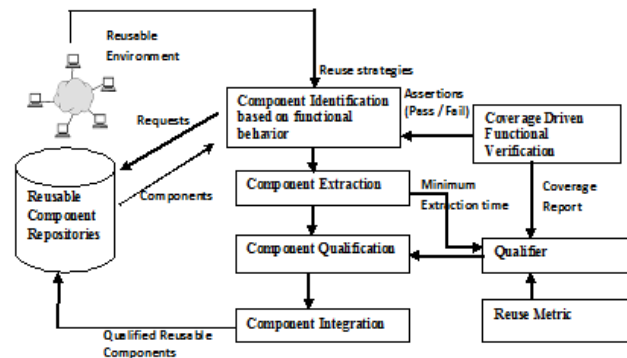


Fig. 1 Conceptual Reference Framework for Reusable Software Components

The whole process consists of the following phases

1. Component identification
2. Component extraction
3. Component Qualification

We focus on a problem as how to analyze the existing replicant and faulty component for identifying the collection of suitable components for use. After once they identified how it is extracted and how to qualify them for appropriate use is the next issue. Our approach to identification, extraction and qualification of reusable software components is based on functional behavior of the component, minimum component extraction time and reusability metrics.

In the component identification phase, to identify the component the information about the components must be encoded. In our approach both replication of software components and the failover components are maintained in the repositories. The functionality of the component is indexed and the functionality request from the existing environment will be compared with the indexes and a collection of match components will be extracted. So the Reuse Utility Percent and Reusable Frequency can be improved by the reuse of replication and faulty components. Coverage driven functional verification method generates the test case for functional verification of the components and the reused components were identified.

Component Identification process is divided into two steps.

- (1) The existing environment sends the request based on functionality.
- (2) The coverage driven functional verification is done for both replicant components and as well as faulty behavior components. And a set of components that possibly satisfying the need is returned.

The next phase in this framework is component extraction. In this phase the identified components which are scattered in the existing environment is extracted for reusability. In our approach we propose a new Component Extraction scheme named as Minimum Extraction Time First (METF).

In this proposed scheme we extract a component which is in the nearest distance to the current position of the environment. By using this scheme we can extract the entire identified reusable component in average minimum extraction time from the current position in the existing environment. This scheme defines an optimal path for component extraction and also calculates extraction time for each component.

After extracting a set of reusable software components, in the qualification phase we qualify the reusable software components based on the coverage report, minimum extraction time to extract the components and software reusability metrics. After qualification the component is integrated with existing environment so that the system is able to continue its usual work.

The following general algorithm illustrates the software component reusability process proposed in this framework

**Begin**

- Step 1: Specify the Reuse Strategy
- Step 2: Index the components with coverage driven functional verification Report
- Step 3: **If** identical match with Components in repositories or Components with faulty behavior **then** identify the component for reuse
- Step 4: **For** entire set of identified component for reuse **do**  
 Calculate the minimum extraction time with METF scheme
- Step 5: Qualify the component based on Coverage report, Minimum extraction time and quality metrics
- Step 6: Integrate the reusable software component with the environment

**End**

**4. Metrics to assess reusability**

A pair of metrics Reuse-Utility-Percent and Reuse Frequency has been used to measure the reusability of the components. Linguistic variables are then assigned to the metrics based on the measurement. The assignment of the linguistic variables depends on the range of the values of the measurement.

Reuse-Utility-Percent is the most important reuse metrics and this metric is very simple to measure. It is the ratio of No. of Reused software components to the No. of software components available in the existing environment. Reuse-Utility-Percent is assigned with six linguistic variables VERY HIGH, HIGH, MEDIUM, LOW, VERY-LOW and NIL as constants in the range of 0-100 in Table 1.

$$\text{Reuse UtilityPercent} = \frac{n(RSC)}{n(SC)} * 100 \quad (1)$$

where n(RSC) is total number of Reusable Software Components & n(SC) is total number of Standard Components existing environment

Table 1 : Linguistic variables for Reuse Utility Percent

<i>Reuse Frequency Range</i>	<i>Linguistic variables</i>
0 – 10	NIL
10 – 30	VERY LOW
30 – 50	LOW
50 – 70	MEDIUM
70 - 90	HIGH
90 - 100	VERY HIGH

Reuse-Frequency is the ratio of the count of a component referred for reuse to the total count of references of the entire standard component in the existing environment.

Reuse-Frequency is assigned to two linguistic variables LOW and HIGH as constants in the range of less than 1 and greater than 1.

$$\text{Reuse Frequency} = \frac{n(C)}{\frac{1}{n} \sum_{i=1}^n n(Si)} \quad (2)$$

where n(C) is total number of reference to a Reusable Software Component, n(Si) is total number of reference

for each Standard Components in the existing environment & n is the total number of component in the existing environment in the Table 2

Table 2: Linguistic variables for Reuse-Utility-Ratio

<i>Reuse Utility Percent Range</i>	<i>Linguistic variables</i>
<=1	LOW
>1	HIGH

The equation (2) shows that the Reuse Frequency is the measure of function usefulness of a component. Hence there should be some minimum value of Reuse Frequency to make software component really reusable.

### 5. Case Study and Analysis

In this section we present a case study to evaluate the effectiveness of component extraction scheme and analysis of reusability metrics with our own test cases

#### 5.1 Evaluation and Discussion

The minimum extraction time for all the components in the entire identified set is calculated. The Total Component Extraction time and Average Component Extraction for the proposed scheme is derived in this subsection. Total Component Extraction time is the total time taken to extract all the reusable components which are identified based on the functionality in Component Identification Phase. Average Component Extraction time is the average time taken to extract all the reusable components

For experimental study Local Area Network Environment with following specification where chosen:

- No. of Nodes=5000 i. e node 0 to node 4999 i. e Distance[C<sub>end</sub>]=4999
- Present reuser position is 143 i. e distance [C<sub>st</sub>]=143 rd node
- The Component Extraction distances of [C<sub>i</sub>] were chosen in Table 3
- C<sub>i</sub> denotes the position in the Network from where the Component for the i<sup>th</sup> position has to be extracted.

Table 3: Distances of C<sub>i</sub> for different Reusable Components of 'i'

<i>Identified Component i</i>	<i>Distance of C<sub>i</sub></i>
-------------------------------	----------------------------------

1	86
2	1470
3	913
4	1774
5	948
6	1509
7	1022
8	1750
9	130

The component Extraction path and minimum extraction time each component in the optimal path for proposed scheme shown in the Table 4.

Table 4: Extraction path & Extraction time using METF

<i>Extraction Path</i>	<i>Distance of C<sub>i</sub></i>	<i>Extraction Time of component C<sub>i</sub></i>
9	130	0.004
1	86	0.016
3	913	0.246
5	948	0.255
7	1022	0.276
2	1470	0.400
6	1509	0.411
8	1750	0.478
4	1774	0.485

The Total distance travelled is,  $T_D = | \text{Distance } [C_{st}] - \text{Distance } [C_9] | + | \text{Distance } [C_9] - \text{Distance } [C_1] | + | \text{Distance } [C_3] - \text{Distance } [C_1] | + | \text{Distance } [C_5] - \text{Distance } [C_3] | + | \text{Distance } [C_7] - \text{Distance } [C_5] | + | \text{Distance } [C_2] - \text{Distance } [C_7] | + | \text{Distance } [C_6] - \text{Distance } [C_2] | + | \text{Distance } [C_8] - \text{Distance } [C_6] | + | \text{Distance } [C_4] - \text{Distance } [C_8] |$

$$T_D = 13+44+745+827+35+74+448+39+241+24 = 1745$$

$$\text{Average distance} = 1745/9 = 193.88$$

Total extraction time = 2.570  
 Average extraction time = 0.286

Extraction path for proposed Extraction Minimum Extraction Time First (METF) method is given in the below Fig 2.

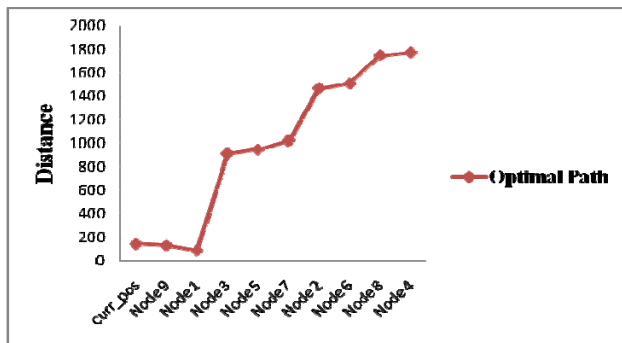


Fig. 2 Extraction path of Reusable Component Extraction (METF)

### 5. 2 Analysis of reusability metrics

For a better explanation of reusability metrics, consider the example of a Local Area Network Environment consists of N=500 components and for 10 scenarios in various cases, the Reuse Utility Percent metric is calculated by using the equation (1) and the linguistic variables are assigned for the range of metric value which is already assigned in the Table 1

Total No. of Standard Components n(SC) = 500

Table 5: Linguistic variables for Reuse-Utility-Percent for different Test cases from NIL to VERY HIGH

Test Cases	No. of Reused Components n(RSC)	Reuse Utility Percent	Linguistic variables
1	59	11.8	VERY LOW
2	385	77	HIGH
3	193	38.6	LOW
4	412	82.4	HIGH
5	323	64.6	MEDIUM
6	215	43	LOW
7	89	17.8	VERY LOW
8	5	1	NIL
9	455	98	VERY HIGH
10	299	59.8	MEDIUM

Table 5, summarizes the number of test cases collected under the LAN environment and Reuse Utility Percent is calculated using the equation (1) and linguistic variables are assigned. From this table, it is inferred that the reuse-utility percent increases when number of component reused is increases

From the below graph in Fig.3 we can infer the relationship between the No. of Reused Components and Reuse-Utility-Percent as Reuse-Utility-Percent is directly proportional to the no. of Reused Components.

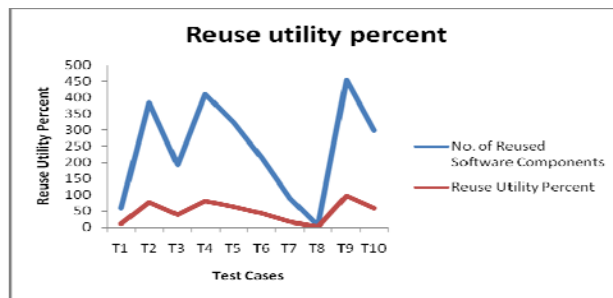


Fig. 3 Analysis Graph for Reuse Utility Percent metric

Consider an another example application of a local area network consists of N=10 reusable components in an application for various references, the reuse frequency metric is calculated and the linguistic variables are assigned for the range of metric value which is already assigned in the Table 2

Total No. of reference for Standard Components  $\Sigma n(S_i) = 225$

Table 6: Linguistic variables for Reuse-Frequency for the reused Components

Component Number	No. of Reference to a Component n(C)	Reuse Frequency	Linguistic variables
1	8	0.35	LOW
2	0	0	LOW
3	15	0.6	LOW
4	54	2.4	HIGH
5	23	1	LOW
6	33	1.46	HIGH
7	4	0.1	LOW
8	42	1.86	HIGH
9	19	0.84	LOW
10	27	1.2	HIGH

Table 6, summarizes the number of components reused under the LAN environment and reuse-frequency is calculated using the equation (2) and linguistic variables are assigned. From this table, it is inferred that the reuse-frequency for a component increases when number of references for a component is increases

From the below graph in Fig.4 we can infer the relationship between No. of references to a Component and Reuse-frequency as Reuse-frequency is directly proportional to the total no. of references to a Component in the existing environment.

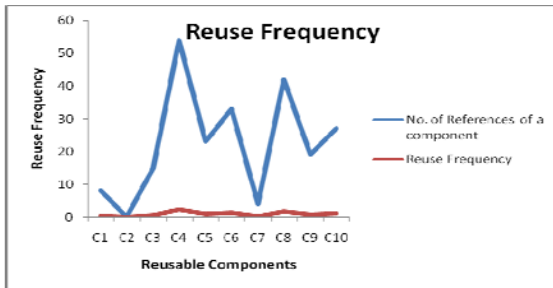


Fig. 4 Analysis Graph for Reuse Frequency metric

## 6. Conclusions

In this paper, we described a framework for reusing the software components in context level. This framework also provides diversity in the execution environment leading to a higher level of reliability of the system. By this new approach we can include the reuse benefits as reduction in development effort and maintenance effort. This approach is feasible for building reliable software systems using the reusable components. Finally, the complete cycle of phases can enable the reuser to determine which components have high reuse potential with regards to specific functional requirements, minimum extraction time and measures of reusability metric. This research also shows how metrics can be used to find the quality attributes of a software component. The long term goal of this approach is to provide a distributed software component repository to support the system development through internet and cloud services.

## References

- [1] Matteo Gaeta, Francesco Orciuoli, Stefano Paolozzi, and Saverio Salernol, "Ontology Extraction for Knowledge Reuse: The e-Learning Perspective", IEEE Transactions on Systems, Man and Cybernetics, Vol. 41, No. 4, pp. 789-809 Jul 2011.
- [2] Gan Wang, Ricardo Valer and Jared Fortune, "Reuse in Systems Engineering IEEE Systems Journal, Vol. 4, No. 3, pp. 376-384, Sep. 2010
- [3] Hokamura, K., Ubayashi, N., Nakajima, S., Iwai, A, "Reusable aspect components for Web applications" in TENCON 2010 - 2010 IEEE Region Conference , pp. 1059 – 1064, Nov. 2010
- [4] Freya H. Lin, Timothy K. Shih, , and Won Kim, "An Implementation of the CORDRA Architecture Enhanced for Systematic Reuse of Learning Objects", IEEE Transactions on Knowledge and Data Engineering , Vol. 21, No. 6, pp 925 – 938, JUNE 2009
- [5] Parvinder Singh Sandhu and Hardeep Singh, "A Fuzzy-Inference System Based Approach for the Prediction of Quality of Reusable Software Components", International Conference on

Advanced Computing and Communications, ADCOM 2008. On page(s): 349 – 352

[6] Parvinder Singh Sandhu and Hardeep Singh, "Automatic Reusability Appraisal of Software Components using Neuro-Fuzzy Approach", International Journal Of Information Technology, vol. 3, no. 3, 2006, pp. 209-214.

[7] Richard W. Selby , "Enabling Reuse-Based Software Development of Large-Scale Systems," IEEE Transaction of Software Engineering, Vol. 31, No. 6, PP. 495-510, Jun 2005

[8] Tomer, L. Goldin, T. Kuflik, E. Kimchi, and S.R. Schach, "Evaluating Software Reuse Alternatives: A Model and its Application to an Industrial Case Study," IEEE Trans. Software Eng., vol. 30, no. 9, pp. 601-612, Sept. 2004

[9] Marcus A. Rothnberger, Kevin j. Dooleg, Uday R. Kulkarni and Nader Nada, " Strategies of Software Reuse : A Principal Component Analysis of Reuse Practices", IEEE Trans. Software Eng., vol. 29, no. 9, pp. 825-837, Sep. 2003.

[10] Morel. B and Alexander, "A Slicing Approach for Parallel Component Adaptation," Proc. 10th IEEE International Conference and Workshop the Eng. of Computer-Based Systems, pp. 108-114, Apr. 2003

[11] Vitharana. P, Zahdi. F and Jain. H, "Design, retrieval and assembly in component-based software development", Communications of the ACM, 46(11), 97-102, 2003

[12] Guo. J, Luqui, "A Survey of Software Reuse Repositories", 7th IEEE International Conference and Workshop on the Engineering of Computer Based Systems, pp. 92-100, Apr. 2000.



# Integrated three-dimensional reconstruction using reflectance fields

Maria-Luisa Rosas<sup>1</sup> and Miguel-Octavio Arias<sup>2</sup>

<sup>1,2</sup> Computer Science Department, National Institute of Astrophysics, Optics and Electronics,  
Puebla, 72840, Mexico

## Abstract

A method to obtain three-dimensional data of real-world objects by integrating their material properties is presented. The material properties are defined by capturing the Reflectance Fields of the real-world objects. It is shown, unlike conventional reconstruction methods, the method is able to use the reflectance information to recover surface depth for objects having a non-Lambertian surface reflectance. It is, for recovering 3D data of objects exhibiting an anisotropic BRDF with an error less than 0.3%.

**Keywords:** *Three-dimensional reconstruction, Reflectance fields, Computer Vision, Computer Graphics.*

## 1. Introduction

Many of the current methods in 3D computer vision rely on the assumption that the objects in the scene have Lambertian reflectance surface (such property is related to the materials that reflect the same amount of incident energy illumination uniformly over all the surface). Unfortunately, this assumption is violated for almost all real world objects, leading to incorrect depth estimates [1][2].

In the area of computer graphics, the reflection from surfaces is typically described by high dimensional reflectance functions. However, the formulation of analytical models is not always an easy task. An alternative approach to the specification of the reflectance or optical properties of the surface objects by analytical modeling is the capture of this reflectance information from real-world surfaces. The acquisition is carried out, with a camera or array of cameras to obtain a set of data that describes the transfer of energy between a light field of incoming rays (the illumination) and a light field of outgoing rays (the view). Such set of data is known as the Reflectance Field [3].

This document explores the problem of obtaining the three-dimensional reconstruction of objects exhibiting an

anisotropic BRDF (the objects material have the property that their reflection characteristics vary to rotations of the surface about its normal) by using a 4D slice of the 8D reflectance field. The 4D slice of the reflectance field is obtained by a camera-projector pair. Our method exploits the property of reciprocity of the reflectance field to impose the epipolar constraint by considering the camera-projector pair as a stereo system. As an example, we show how our method can be used to recover objects with an anisotropic BRDF of their surface. This procedure avoids the need of an analytical model of the reflectance data.

## 2. Theory

### 2.1 Reflectance field and Light transport constancy

Debevec et al [3] showed that the exiting light field from the scene under every possible incident field of illumination can be represented as an 8D function called the reflectance field:

$$R(L_i(\psi_i); L_0(\psi_0)) = R(\psi_i; \psi_0) \quad (1)$$

Here,  $L_i(\psi_i)$  represents the incident light field on the scene, and  $L_0(\psi_0)$  represents the exiting light field reflected off the scene. In order to work with discrete forms of these functions, the domain  $\psi$  of all incoming directions can be parameterized by an array indexed by  $i$ . The outgoing direction corresponding to an incoming direction is also parameterized by the same index,  $i$ . Now, consider emitting unit radiance along ray  $i$  towards the scene (e.g., using a projector). The resulting light field, which is denoted by vector  $\mathbf{t}_i$ , captures the full transport of light in response to this impulse illumination. This is called the impulse response or the impulse scatter function [4].

All the impulse responses can be concatenated into a matrix  $\mathbf{T}$  which is called the light transport matrix:

$$\mathbf{T} = [\mathbf{t}_1 \mathbf{t}_2 \dots \mathbf{t}_n]$$

(2)

Since light transport is linear, any outgoing light field represented by a vector  $\mathbf{L}_0$  can be described as linear combination of the impulse responses,  $\mathbf{t}_i$ . Thus, for an incoming illumination described by vector  $\mathbf{L}_i$ , the outgoing light field can be expressed as:

$$\mathbf{L}_0 = \mathbf{T}\mathbf{L}_i \quad (3)$$

The light transport matrix  $\mathbf{T}$ , is thus the discrete analog of the reflectance field  $R(L_i(\psi_i); L_0(\psi_0))$ .

## 2.2 Symmetry of the transport matrix

The idea that the flow of light can be effectively reversed without altering its transport properties was proposed by von Helmholtz in his original treatise in 1856 [5]. He proposed the following reciprocity principle for beams traveling through an optical system (i.e., collections of mirrors, lenses, prisms, etc.):

*Suppose that a beam of light  $\mathbf{A}$  undergoes any number of reflections or refractions, eventually giving rise (among others) to a beam  $\mathbf{B}$  whose power is a fraction  $\mathbf{f}$  of beam  $\mathbf{A}$ . Then on reversing the path of the light, an incident ray  $\hat{\mathbf{B}}$  will give rise to a beam  $\hat{\mathbf{A}}$  whose power is the same fraction  $\mathbf{f}$  of beam  $\hat{\mathbf{B}}$ .*

In other words, the path of a light beam is always reversible, and furthermore the relative power loss is the same for the propagation in both directions. For the purpose of a reflectance field generation, this reciprocity can be used to derive an equation describing the symmetry of the radiance transfer between incoming and outgoing directions  $\psi_i$  and  $\psi_0$ :

$$R(\psi_i; \psi_0) = R(\psi_0; \psi_i) \quad (4)$$

where  $R$  is the reflectance field. For the light transport matrix defined in the last section, this implies that the transport of light between a ray  $i$  and a ray  $j$  is equal in both directions, i.e.

$$T[i, j] = T[j, i] \implies T = T^T \quad (5)$$

Therefore,  $T$  is a symmetric matrix(See work in [6]).

## 2.2 BRDF

The Bidirectional Reflectance Distribution Function (BRDF) is a projection of the 8D reflectance field into a lower dimension. From equation 1, the 4D reflectance field can be represented as

$$f_r(L_i(\Omega_1); L_0(\Omega_2)) = f_r(\Omega_1; \Omega_2) \quad (6)$$

where  $L_i(\Omega_1)$  represents the incident light field on the scene, and  $L_0(\omega_2)$  represents the exiting light field reflected off the scene and  $\Omega_1, \Omega_2$  are incoming and outgoing directions, e.g.,  $(\theta_1, \phi_1), (\theta_2, \phi_2)$ . In essence, the BRDF describe how bright the differential surface  $dA$  of a material appears when it is observed from a certain direction and illuminated from a certain direction.

The reciprocity exposed in the last section, the 4D reflectance field can be written as

$$f_r(\Omega_1; \Omega_2) = f_r(\Omega_2; \Omega_1) \quad (7)$$

Some materials have the property that their reflection characteristics are invariant to rotations of the surface about its normal. Such materials are called isotropic. Materials not having this characteristic are called anisotropic. As equation 2 shows, in order to discretize the equation 7, all incoming and outgoing directions in domain  $\Omega$  can be parameterized by an array indexed by  $i$ . We denote the resulting 4D light field by vector  $\hat{\mathbf{t}}_i$  and this 4D light field is concatenated as

$$\hat{\mathbf{T}} = [\mathbf{t}'_1 \mathbf{t}'_2 \dots \mathbf{t}'_n] \quad (8)$$

For an incoming illumination described by vector  $\mathbf{L}'_i$ , the outgoing light field can be expressed as

$$\mathbf{L}'_0 = \hat{\mathbf{T}}\mathbf{L}'_i \quad (9)$$

The matrix  $\hat{\mathbf{T}}$  is the discrete analog of the 4D reflectance field. The reciprocity exposed in the last section implies that the transport of light between a ray  $i$  and a ray  $j$  is equal in both directions, i.e.

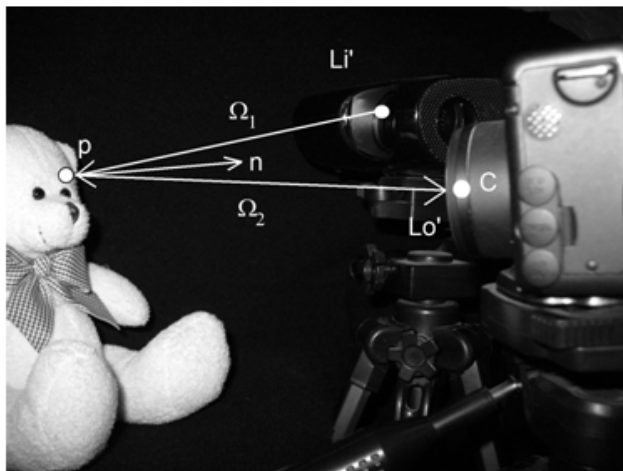
$$\hat{T}[i, j] = \hat{T}[j, i] \implies \hat{T} = \hat{T}^T \quad (10)$$

## 3. Depth recovery from the 4D reflectance field

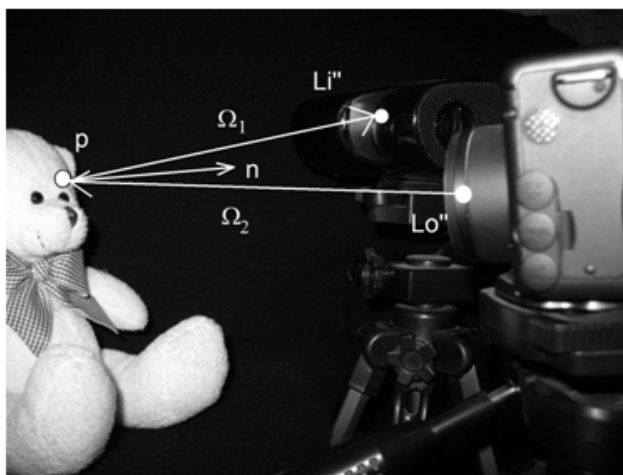
Consider the scene configuration in Fig. 1a. All the scene is illuminated by a projector  $\mathbf{L}'_i$ . A particular point in the scene  $p$  will reflect light to the camera  $C$  according to equation 9, the outgoing light field represented by the vector  $\mathbf{L}'_0$  is the reflected intensity in the direction of  $C$  from the point  $p$  with normal vector  $\hat{n}$ . Let  $o_1$  and  $o_2$  denote the positions of the projector and camera, respectively. The unit vectors  $\Omega_1 = \frac{1}{|o_1-p|}(o_1-p)$  and  $\Omega_2 = \frac{1}{|o_2-p|}(o_2-p)$  denote the directions from  $p$  to the projector and camera, respectively. Given this configuration, the image irradiance (see [7]) at the projection of  $p$  is

$$e = f_r(\Omega_2, \Omega_1) \frac{\hat{n} \cdot \Omega_2}{|\Omega_2 - p|^2} \quad (11)$$

where  $f_r$  is the BRDF (4D function).



a.



b.

Fig. 1 The scene in (a) is illuminated by a light source  $L_i'$ . A particular point in the scene  $p$  will reflect light to the camera C. The outgoing light field  $L_o'$  is the reflected intensity in the direction of C from the point  $p$ . The scene in (b) is “illuminated” by a camera  $L_o''$ . A particular point in the scene  $p$  will reflect light and it is “captured” by a light source. The outgoing light field  $L_i''$  is the reflected intensity in the direction of light source from the point  $p$ .

In the above equation is assumed that every ray light  $\Omega_2$  from the light source illuminates the scene and the number of rays reflected is just one, it can be considered true for those objects with 4D material properties. Now we add the transport matrix  $\hat{T}$  to the equation 11, so we have,

$$e = \hat{T}(p) \frac{\hat{n} \cdot \Omega_2}{|\Omega_2 - p|^2}$$

where  $\hat{T}(p)$  is the 4D transport matrix that corresponds to a point  $p$  of the scene,  $\hat{n}$  can be expressed as  $(\frac{dz}{dx}, \frac{dz}{dy}, -1)$ , the ray from the camera can be expressed as  $\Omega_2(p) = (\Omega_{2x}, \Omega_{2y}, \Omega_{2z})$

Taking advantage of the symmetry of the transport matrix we can impose the epipolar constraint to provide a solution to equation 12. Consider the scene configuration in Fig. 1b. All the scene is “illuminated” by a camera  $L_o''$ . A particular point in the scene  $p$  will reflect light and it is “captured” by a light source. Then, we can consider the system configuration as a stereo setup such as, it can be calibrated as a stereo system.

The vector  $\Omega_2(p)$  and the denominator  $|\Omega_2 - p|^2$  can be determined when calibrating a stereo setup. Imposing the epipolar constraint we can express the normal  $\hat{n}$  as  $(\frac{dz}{dx}, 0, -1)$ .

The point  $p(x, y, z)$  will have projections in the camera and the light source (considered as a second camera) established by calibration parameters of the system. Expressing the depth as  $z(x, y)$ , we rewrite the equation 12 as

$$\frac{dz}{dx} = \frac{e|\Omega_r - p|^2 \hat{T}(p) + \Omega_2 z}{\hat{T}(p) \Omega_{2x}} \quad (13)$$

This can be numerically integrated as

$$z(x, y) = \int_{x_0}^x \frac{dz}{dx} dx + z(x_0, y) \quad (14)$$

For each epipolar line  $y$ , this integral provides the depth across the epipolar line. We can determine for each epipolar line  $y$  the  $z(x_0, y)$  since the point  $p(x, y, z)$  have projections in the camera and we know the corresponding projections to the light source when the 4D transport matrix is captured.

### 3. Test reconstruction

In order to obtain the three-dimensional reconstruction of the object placed in the scene some calibration parameters have to be computed, to do that we use the Dual Photography [6] technique to use the camera-projector assembly as a stereo system for enabling the projector to “capture” images like a camera, thus making the calibration of a projector essentially the same as that of a camera, which is well established. A standard black-and-white checkerboard is used. The flat checkerboard positioned with different poses is imaged by the camera

and poses from the point of view of the projector are generated. Once, these poses are obtained the intrinsic and extrinsic parameters of the stereo system using the Matlab toolbox provided by Bouguet [8] are computed.

Fig. 2 shows an example of the checkerboard images captured from the point of view of the camera (a) and synthesized from the point of view of the projector (b).

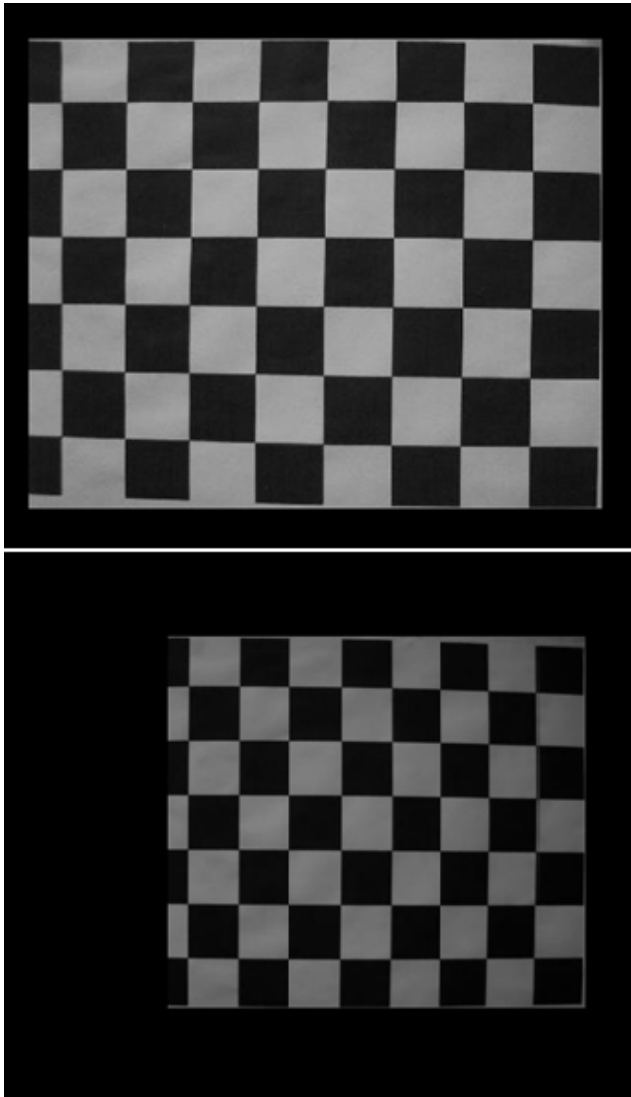
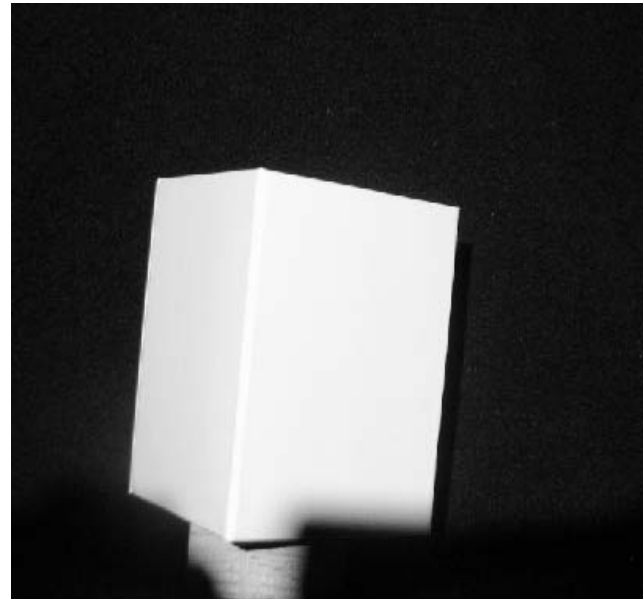


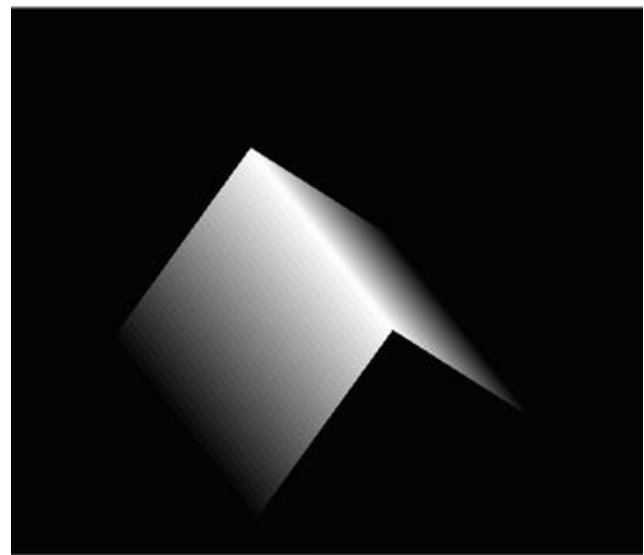
Fig. 2 Images synthesized of the checkerboard from the point of view of the camera (top) and from the point of view of the projector (bottom).

The three-dimensional reconstruction using 4D light field was implemented in Matlab and validated its effectiveness in the following experiment. We obtained the reconstruction of a real object exhibiting a non-Lambertian material which dimensions are known and the RMS error was computed by comparing the reconstructed object and the real object measurements. The Fig. 3 shows the three-dimensional reconstruction of a cube. The RMS

error between the cube recovered and the real cube is of 0.3%.



a.



b.

Fig. 3 Three-dimensional reconstruction (b) of a real object exhibiting a non-Lambertian material from its surface (a).

#### 4. Conclusions

All methods of three-dimensional reconstruction in computer vision area are influenced by light and the material properties of the objects. The estimation of the material of reflectance properties of the object is important for a correct 3D measurement. In computer graphics, the

material properties of such objects materials are measured and they are described as a dimensional reflectance functions (8D function). The theory and experiment have demonstrated the ability of obtaining the three-dimensional reconstruction of objects exhibiting an anisotropic BRDF by integrating a 4D slice of the 8D reflectance field information by using a camera-projector pair with an error less than 0.3% of the real-world object measurements. Also, this procedure represents the first step to extend the formulation to include 6D and 8D surface properties.

### Acknowledgments

The authors acknowledge the partial support by Mexican Conacyt.

### References

- [1] C. Yannick., SPINLER K., BOURENNANE S., and WITTENBERG T.: 'New structured illumination technique for the inspection of high-reflective surfaces: application for the detection of structural defects without any calibration procedures'. *J. Image Video Process. 2008* (2008), pp. 1–14
- [2] ANGELO P., and WOHLER C.: '3D surface reconstruction based on combined analysis of reflectance and polarisation properties'. In *Society of Photo-Optical Instrumentation Engineers(SPIE) Conference Series* (June 2005), Osten W., Gorecki C., Novak E. L., (Eds.), vol. 5856 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pp. 491–502
- [3] DEBEVEC P., HAWKINS T., TCHOU C., DUIKER H.P., SAROKIN W., and SAGAR M.: 'Acquiring the reflectance field of a human face'. In *Siggraph 2000, Computer Graphics Proceedings* (2000), Akeley K., (Ed.), *ACMPress/ACMSIGGRAPH/Addison Wesley Longman*, pp. 145-156
- [4] SEN P., CHEN B., GARG G., MARSCHNER S., HOROWITZ M., LEVOY M., and LENSCH H.: 'Dual Photography'. *ACM Transactions on Graphics* 24, 2005, pp. 745-755
- [5] ZICKLER T., BELHUMEUR P., and KRIEGMAN D.: 'Toward a stratification of helmholtz stereopsis'. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on* 1, 2003, 548
- [6] BOUGUET J.: 'Camera calibration toolbox for matlab'. <http://www.vision.caltech.edu>

**First Author** obtained her B.Eng. in Computer Science at the UPAEP (University of Puebla) in Puebla, Mexico, in 2002. She obtained a M.Sc. in Computer Science at the INAOE (National Institute of Astrophysics, Optics and Electronics, Puebla, Mexico) in 2004. For two years (2006-2008), she worked in Prefixa Vision Systems (Puebla, Mexico) where she developed a 3D Camera. Since 2008 she is a Ph.D student in the Computer Science department at the INAOE. She is an inventor of the patent: Method and apparatus for rapid three-dimensional restoration. Her current research interests are computer vision, computer graphics, FPGA and CUDA architectures, robotics and genetic algorithms.

**Second Author** obtained his B.Eng. in Communications and Electronics at the FIMEE (University of Guanajuato) in Salamanca, Gto. in 1990. He also obtained a M.Eng. in Instrumentation and Digital Systems at the FIMEE two years later. In 1997, he finished his Ph.D. degree at the Computer Vision and Systems Laboratory of Université Laval (Quebec city, Canada). He was a professor-researcher at the Computer and Systems Laboratory at Laval University where he worked on the development of a Smart Vision Camera. Since 1998 he is with the Computer Science department of INAOE (National Institute of Astrophysics, Optics and Electronics, Puebla, Mexico) where he continues his research on FPGA architectures for computer vision.

# Reasoning in Graph-based Clausal Form Logic

Alena Lukasova<sup>1</sup>, Martin Zacek<sup>2</sup>, Marek Vajgl<sup>3</sup>

<sup>1</sup> Department of Informatics and Computers, University of Ostrava  
Ostrava, Czech Republic

<sup>2</sup> Department of Informatics and Computers, University of Ostrava  
Ostrava, Czech Republic

<sup>3</sup> Department of Informatics and Computers, University of Ostrava  
Ostrava, Czech Republic

## Abstract

This paper follows the work of T. Richards specialization Clausal Form Logic formal system of the first order logic. The paper presents also the way of using graph-based clausal form statements in the frame of semantic (associative) networks. The goal of our research is to follow the direction towards graph-based clausal form knowledge representation shaped by Richards and build up a graph-based formal system. The new formal system Graph-based Clausal Form Logic has its own graph-based language with the expressivity similar to that one of CFL. The idea of the GCFL graph-based approach is also useful in the frame of the RDF model especially in its graph version.

**Keywords:** *Clausal Form Logic, formal system, graph, Graph-based Clausal Form Logic, RDF model.*

## 1. Introduction

T. Richards in his CFL (Clausal Form Logic) formal system of the first order logic (FOL) follows the idea of the “Horn’s clauses” and generalizes it to a concept of “conditional clauses”, especially useful in declarative programming languages like PROLOG.

He defines CFL as a formal system with a special language with model-theoretic semantics and introduces a CFL modification of resolution inference rule as a tool of a formal reasoning.

He also shows the way of using graph-based clausal form statements in the frame of semantic (associative) networks. Richards [1] uses the graph-based representation especially for the illustration of meaning of clausal form expressions of CFL.

The goal of our approach is to follow the direction towards graph-based clausal form knowledge representation shaped by Richards, and build up a graph-based formal system that does not only graphically illustrate knowledge bases, but also allows users to obtain consequents of a knowledge base in a graph-based way.

The new formal system GCFL (Graph-based Clausal Form Logic) has its own graph-based language with the expressivity similar to that one of the CFL; moreover it uses the inference methods of associative networks [2] and proposes a graph-based modification of the resolution inference method of reasoning corresponding to that one of the CFL.

The idea of the GCFL graph-based approach is also useful in the frame of the RDF model especially within its graph version. To create RDF(S) knowledge base using the GCFL language completed by corresponding URIs is not difficult comparatively with an approach of OWL language representation. Finally the GCFL serves an own easy-to-understandable and usable inference mechanism.

## 2. Richard’s clausal form logic (CFL) and its graph-based modification (GCFL)

As one of the main tools of formal reasoning, the CFL introduced by T. Richards [1] uses the conditional „if – then“ statements.

A simple example illustrates the “if – then“ statement structure by a “Holmes rule”:

*“If one person  $x$  hates another person  $y$ , then  $x$  knows  $y$ .”*

or

*If  $hate(x,y)$  then  $know(x,y)$*

or

*If  $\langle antecedent \rangle$  then  $\langle consequent \rangle$*

Generally a conditional statement (clause) proposed by T. Richards says that some (composed) consequent statement follows from another (composed) antecedent statement.

Richards also proposed an alternative representation of the clausal form atoms (vectors) in a graph-based language well known in associative (semantic) networks (Fig. 1).

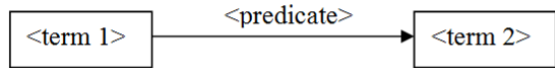


Fig. 1 The associative networks

Our idea of graph based representation of clauses connects to that one of Richards' CFL and proposes a complete graph-based inference system GCFL. Knowledge base of the system consists of clauses represented by graphs.

To distinguish statements in the antecedent part of a graph-based clause from statements in its consequent part we introduce a convention

- to draw the arcs of antecedent vectors by dashed lines and
- to draw the arcs of consequent vectors by solid lines.

All the vectors (with solid or dashed lines) represent atomic statements and have generally the structure

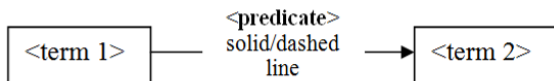


Fig. 2 Vectors

In the Richards' clause language a general formula of CFL is of a form

$$\begin{aligned}
 &\langle \text{antecedent} \rangle \rightarrow \langle \text{consequent} \rangle \\
 &\quad \text{or} \\
 &P_1 \&\dots\& P_m \rightarrow Q_1 \vee \dots \vee Q_n \quad (1) \\
 &\quad \text{or} \\
 &P_1, \dots, P_m \rightarrow Q_1, \dots, Q_n
 \end{aligned}$$

where “ $\rightarrow$ ” is a meta-symbol, the antecedent is a conjunction of some set of positive first order logic atoms (vectors)  $\{P_1, \dots, P_m\}$  and the consequent is a disjunction of another set of positive first order logic atoms (vectors)  $\{Q_1, \dots, Q_n\}$ .

The convention of two mutually different arc lines in the GCFL does not need to separate antecedent and consequent by any meta-symbol like “ $\rightarrow$ ”, all the dashed antecedent vectors are mutually connected by  $\&$  and all the solid consequent vectors are mutually connected by  $\vee$ .

In our Holmes example presented above we draw a vector with dashed-line arc as the antecedent part of graph representing a clause (Fig. 3)

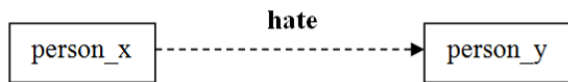


Fig. 3 Holmes rule

and as the consequent part of the graph a vector with solid-line arc (Fig. 4)

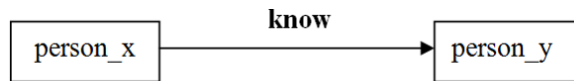


Fig. 4 Vector with solid-line arc

So the Holmes example above has in the graph language GCFL a form a clause network (Fig. 5)

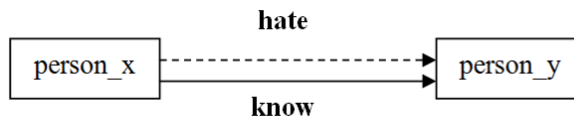


Fig. 5 Clause network

An antecedent or a consequent of the conditional clause in the GCFL can be also an empty set of atoms like in the same way as in the CFL.

In the case of an empty antecedent in the CFL for example the atom

$$\rightarrow \text{know}(\text{person}_x, \text{person}_y)$$

has a meaning of a positive fact “A person<sub>x</sub> knows a person<sub>y</sub>.” To express is using the syntax of the graph language of GCFL a vector like that one at the Fig.4 suffices.

In the case of an empty consequent in the CFL the formal representation is of the form

$$\text{know}(\text{person}_x, \text{person}_y) \rightarrow$$

that represents a negative fact “It is not true that a person<sub>x</sub> knows a person<sub>y</sub>.”

To represent is using GCFL a vector with dashed line arc suffices (Fig.6).

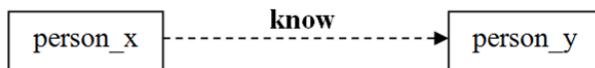


Fig. 6

The structure of the clause allows only constructions of clauses with connections  $\&$  in the antecedent and connection  $\vee$  in the consequent. If necessary GCFL as well as CFL solves the problem of disjunction in the antecedent (conjunction in the consequent) by a following decomposition of the clause into  $m$  ( $n$ ) separate clauses:

$$\begin{array}{ll}
 P_1 \rightarrow Q_1 \vee \dots \vee Q_n & P_1 \&\dots\& P_m \rightarrow Q_1 \\
 \vdots & \vdots \\
 P_m \rightarrow Q_1 \vee \dots \vee Q_n & P_1 \&\dots\& P_m \rightarrow Q_n
 \end{array}$$

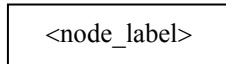
### 3. Abstract syntax of the GCFL

#### Definition (GCFL language symbols)

The language of the formal system GCFL uses the following symbols:

1) Nodes

- a) graphical symbols for nodes of networks

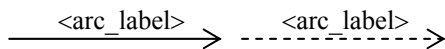


- b) terms in node labels are strings for representation of

- variables – with capitals at first positions: X,Y, Man, Country,...
- constants – with small letters or numbers at first positions: anne, cat, student,...
- function symbols – sin(X), matka(X),...
- existential terms – with @ as a prefix: @anybody, @known(X),...

2) Arcs

- a) graphical symbols for arcs of networks,



- b) binary predicate symbols as arc labels: isa(X,Y), citizen(X,Y),...

#### Definition (vector, ground/universal/existential vector)

Vector (atom) in the GCFL language (Fig. 2) consists of two nodes labelled by <term\_1> and <term\_2> symbols and their dashed/solid connecting arc labelled by a <predicate symbol>.

Vector with only constant labels of its nodes is a ground vector.

Vector with any variable symbol of a node label is a universal vector.

Vector with any existential symbol of a node label is an existential vector.

#### Definition (clause network, knowledge base)

Clause network in the GCFL language is a network of antecedent and/or consequent atomic vectors.

Knowledge base of GCFL is a set of GCFL clause networks.

Negative statement is in the GCFL expressed as a vector with a special symbol  $\otimes$  (see Fig. 7). For example a statement “Anne does not know Jane.” is capture in figure 7 as the vector:

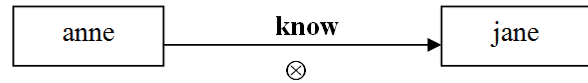


Fig. 7 Negative statement

### 4. Truth and consistency of a conditional clause in a given interpretation I

Truth value of the vector corresponds to the truth value of a corresponding predicate atom in the first order logic interpretation [1].

That means:

A ground vector is *true* iff there is a pair of elements within the relation ordered to the predicate in the *I* that is equal to the pair of ground vector constant terms, otherwise it is *false*.

For universal/existential vector holds that if all its variables are evaluated by constant terms and all of the values of the existential terms are found, then a decision of a true/false interpretation of the vector becomes the same as for the case of a ground vector.

A ground conditional clause is false if all the vectors of the antecedent are true and all the vectors of the consequent are false. Otherwise it is true.

A clause with an empty antecedent is evaluated as true; an empty consequent clause is represented as false.

A conditional clause is *consistent* in a given interpretation *I* if there is a valuation of all the variables that makes the clause true.

### 5. A special role of the predicate isa (ako)

Creating knowledge bases in graph-based formal systems brings (in comparison with the FOL) some new requests. For the sake of expression the clause in GCFL networks all the atoms of first order logic have to be transformed into corresponding binary versions (FOL as well as CFL does not use only binary predicates in its atoms, but also unary, ternary etc. predicates).

For the sake of creating knowledge base in the GCFL

- n-ary predicates (  $n \geq 3$  ) have to be decomposed to binary predicates,
- unary predicates have to use some auxiliary predicates to become corresponding binary predicates with the same meaning.

Ordering of unary predicates into GCFL networks is solved by the usage of special binary predicate symbols **isa**(<term1>, <term2>) with the meaning “is a” or **ako**(<term1>, <term2>) with the meaning “a kind of”.

In the case of our Holmes rule in the CFL language the solution can have a form of a clause



$isa(X, person) \& isa(Y, person) \& hate(X, Y) \rightarrow know(X, Y)$

If the GCFL syntax is applied, then usage of the CFL principles applied on the Holmes rule cause that in the two-part graph (without symbol  $\rightarrow$ ) contains only dashed antecedent vectors and full consequent vectors:

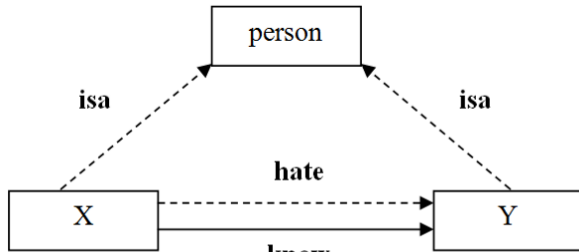


Fig. 8 Holmes rule

Capitals X and Y define variables X and Y, which are universally quantified. Fig. 8 expresses a *universal conditional clause*.

Using FCFL, the existential (skolem) constants are marked with the prefix @ at the beginning of the node name.

For example (see Fig. 9), the existential constants @person\_x and @person\_y introduce the individualities of the two persons.

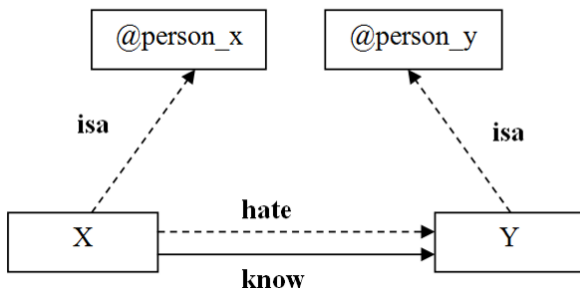


Fig. 9 Networks with existential constants

Predicates **isa** (**ako**) brings some elements of inheritance into a formal system.

If a statement  $r(X, Z)$  with a predicate **r** holds and at the same time  $isa(Y, X)$  holds, then also  $r(Y, Z)$  must hold. So it means in the language of GCFL the following auxiliary rule (Fig. 10) also has to be valid.

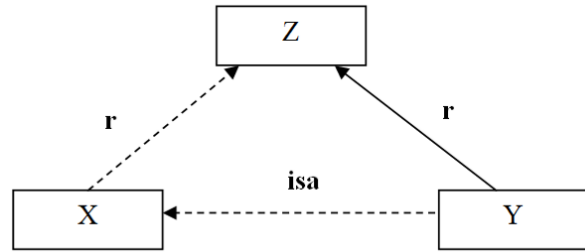


Fig. 10 Auxiliary rule

The clause at the Fig. 11 expresses the statement “If Anne does not know Jane, Jane is a stranger for Anne”

GCFL as well as CFL solves a problem of a base negative vector within a clause by means of “transfer” from antecedent to consequent and vice versa.

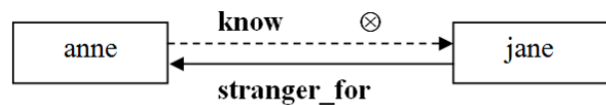


Fig. 11 Statement

For the ground clause holds a following “*rule of the transfer of negative ground atoms*”.

If a negative ground atom (vector) ought to be ordered into the antecedent set of atoms, transfer it as a positive atom into the consequent set of atoms.

If a negative ground atom (vector) ought to be ordered into the consequent set of atoms, transfer it as a positive atom into the antecedent set of atoms.

After a transfer of an antecedent vector into the consequent part the graph (Fig.12) becomes a graph of a clause without antecedent with a meaning “Anne knows Jane or Jane is a stranger to Anne.”

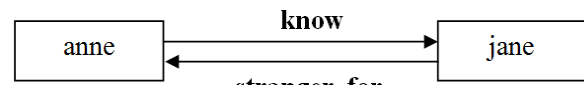


Fig. 12

Similarly GCFL as well as CFL also applies a special transfer rule in the case of universal/existential clause (for example this one of the Fig. 13) in the same way as in the extended RDF graph model, which works with quantifiers [4].

The clause “It is not true that Jane knows everybody, then somebody is a stranger for Jane.” changes after the transfer into a clause “Jane knows everybody or somebody is a stranger for Jane.”(Fig. 14).

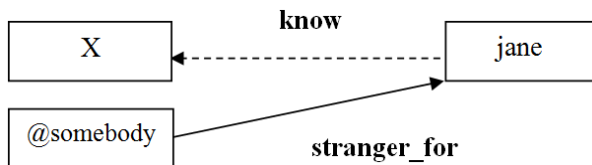


Fig. 13 Universal/existential clause

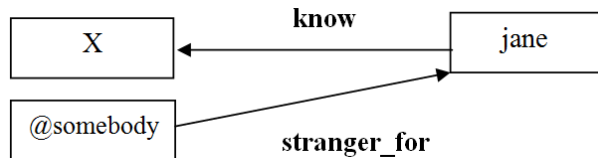


Fig. 14 Result clause

### 6. Resolution reasoning on GCFL knowledge base

GCFL works by the help of two rules - the substitution rule and the cut rule, both are well known in the first order logic. Together both rules form *the resolution rule* as well as in the CFL. In the frame of associative networks [2, 3] the resolution rule has been slightly modified and it is known as the *transfer rule*.

#### The transfer rule

Conditional clause is a tool of transferring its consequent into another clause unfixable with its antecedent. It is possible (for example) to use the *isa* rule (Fig. 8) for transferring its consequent (after proper substitution into another network).

For example, if we apply the vector **have**(rex, fell) to the semantic network N (see Fig. 10), then we get an instance (Fig. 15) describing properties of animals and the position of Rex.

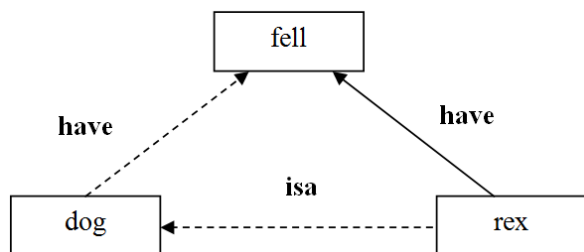


Fig. 15 The transfer rule

#### The substitution rule:

A new clause can be obtained from a clause with variables by a uniform substitution of a term for some of the variables.

#### The cut rule:

If there are two clauses sharing the same atom in the knowledge base for reasoning, one in the antecedent of the

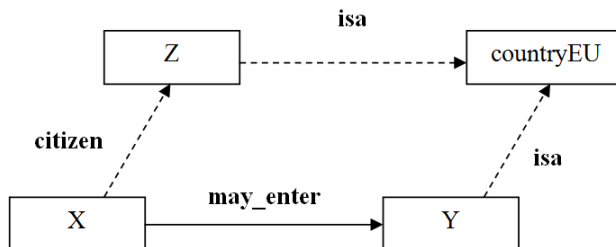
first clause and one in the consequent of the second clause, then we can obtain a new clause by cutting out the same atoms at both sides and create a new clause with antecedent (consequent) that contains all the atoms of the original clauses antecedents (consequents).

### 7. Examples of the resolution reasoning in GCFL

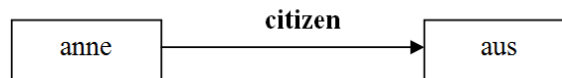
#### Example 1

Immigration rules of an EU country for citizens of EU countries in the language of GCFL forms a knowledge base in the following set of graphs {1, 2, 3, 4}.

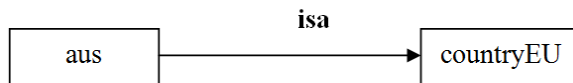
1.



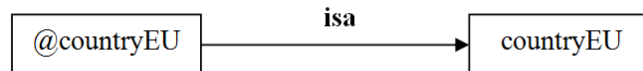
2.



3.

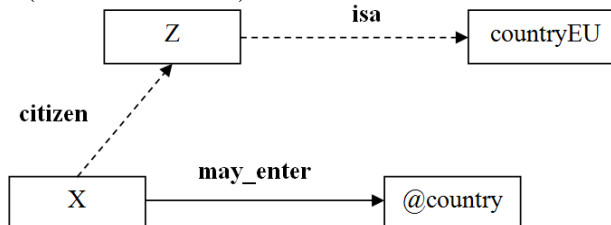


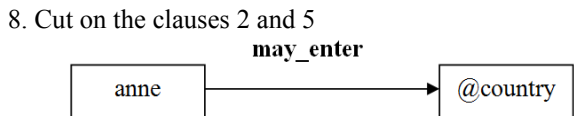
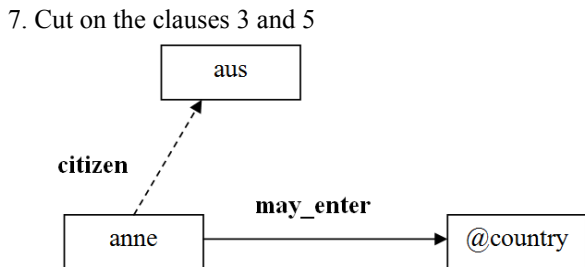
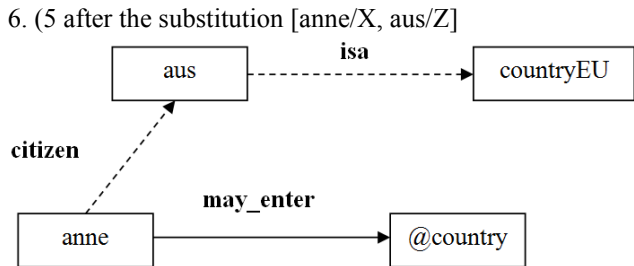
4.



Proof of the statement „Anne may enter.“ in the GCFL on the knowledge base {1, 2, 3, 4}. Graphs 1 – 4 are prerequisites of the proof.

5. (1 after a cut with 4)



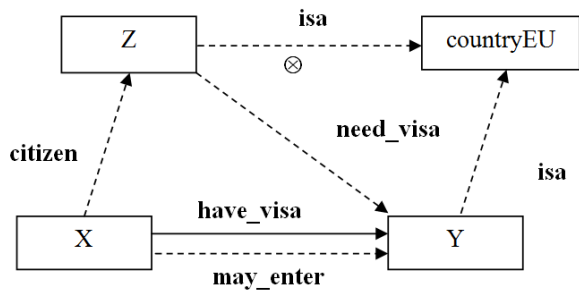


**Example 2**

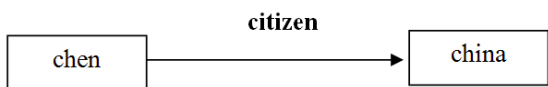
Indirect proof that a citizen of China Chen cannot enter an EU country without a visa:

Graphs 1 – 8 are prerequisites of the proof, graph 9 is a conclusion to be proved. Graphs 10 – 12 are (shortened) steps of the proof.

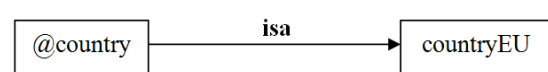
1.



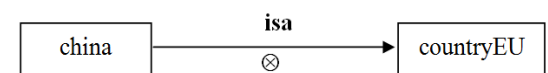
2.



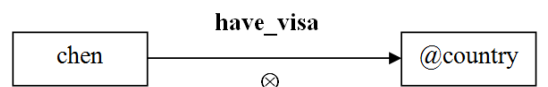
3.



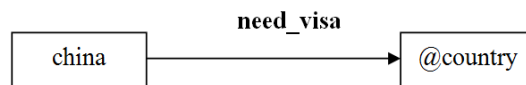
4.



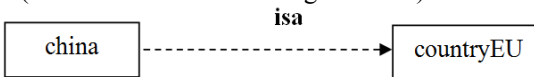
5.



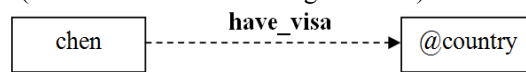
6.



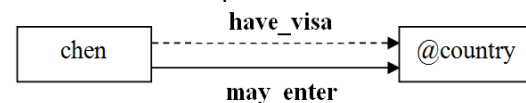
7. (4 after a transfer into a negative fact)



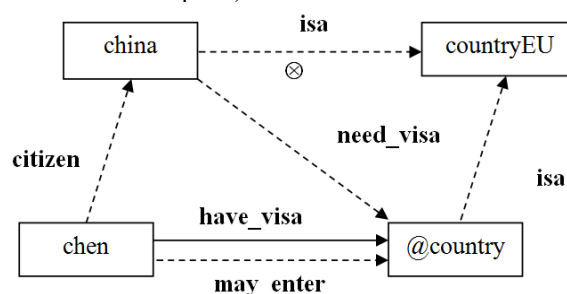
8. (5 after a transfer into a negative fact)



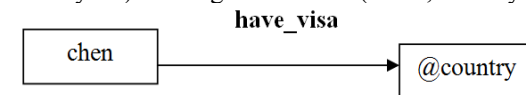
9. A conclusion to be proven



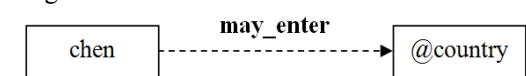
10. (1 after the substitution [chen/X, china/Z, @country/Y] and after the substitution the transfer of negated ground vector isa(china, countryEU) into the clause consequent)



11. (9 after cuts with positive facts citizen(chen, china), need\_visa(china, @country), isa(@country, countryEU) and negative fact isa(china, countryEU))



12. Negation set of clause 9.



13. After two cuts on the graphs (clauses) 11 and 12 the resulting graph becomes empty, it is an inconsistent clause – a contradiction has been obtained.

**8. Conclusions**

As every formal system ought to operate on knowledge bases because of obtaining new consequents and interrelations between them it is natural to develop knowledge systems that can manipulate with structured data in a straightforward graph/based way without a necessity of rewriting knowledge into a language like OWL. Moreover, at present a concept of Linked Data built on RDF model lies at the proper heart of what Semantic

Web is all about. The idea of the RDF knowledge representation in its graph/based modification has its predecessor in a concept of associative (semantic) networks as a simple and understandable tool to represent of knowledge bases.

Our system GCFL gives a possibility to deduce new knowledge or to create new interesting interrelations in a straightforward way within graph representation. This fact offers us a possibility to use notifications and inference rules of GCFL also in the frame of RDF modeling. Reasoning, possibly supported by the graphical version of representation, then becomes more understandable and easier to use than that one of rewriting RDF models by description logic tools (OWL language).

## References

- [1] Richards, T.: Clausal Form Logic. An Introduction to the Logic of Computer Reasoning. Addison-Wesley, 1989.
- [2] Lukasová, A.: Knowledge representation in associative networks (in Czech) Proceedings of Znalosti 2001. Praha, 2001.
- [3] Lukasová, A., Telnarová, Z., Habiballa, H., Vajgl, M.: Formal knowledge representation (in Czech).Universum, 2010. 345 p.
- [4] Lukasová, A., Vajgl, M., Žáček, M.: Reasoning in RDF graphic formal system with quantifiers. Proceedings of the International Multiconference on Computer Science and Information Technology. 2010. pp. 67-72.

**Alena Lukasova** has been working as a professor at the Department of Informatics and Computers at the University of Ostrava (Czech Republic). Her interests include theoretical principles of information and knowledge systems and tools of representation of semantically structured knowledge (database systems with knowledge bases components and their semantics), formal deduction in concept oriented languages, formal ontology for information systems, principles of ontology driven (based) information systems, and sharable knowledge patterns. She is author more than 50 scientific publications.

**Martin Zacek** graduated at the University of Ostrava (Czech Republic). The focus of his dissertation is a formal deduction in graph knowledge representation systems. The PhD thesis includes four graph systems: semantic (associative) networks, conceptual graphs of Sowa, RDF model and Topic Maps. The topic of the PhD thesis corresponds to the content of this article. He is author more than 10 scientific publications. In 2010 he won award "Young Scientist" at the International Multiconference on Computer Science and Information Technology awarded by the International Fuzzy Systems Association Distinction for the presentation of article Reasoning in RDFgraphic formal system with quantifier.

**Marek Vajgl** graduated at the University of Ostrava (Czech Republic) and defended PhD. thesis with title "A proposal of tool for creating and updating knowledge bases for semantic web". He has been working as a lecturer at the Department of Informatics and Computers, University of Ostrava. His interests include descriptive logic formalism which is frequently used as semantic web formalism. He is author more than 20 scientific publications.

# Extraction of Facial Feature Points Using Cumulative Histogram

Sushil Kumar Paul<sup>1</sup>, Mohammad Shorif Uddin<sup>2</sup> and Saida Bouakaz<sup>3</sup>

<sup>1</sup> Department of Computer Science and Engineering (CSE)  
Jahangirnagar University, Savar, Dhaka-1342, Bangladesh  
Phone: + (880) 1711172191

<sup>2</sup> Department of Computer Science and Engineering (CSE)  
Jahangirnagar University, Savar, Dhaka-1342, Bangladesh  
Phone: +(880) 1552471751

<sup>3</sup> Head of the SAARA Research Team, LIRIS Lab, Nautibus Building  
University Claude Bernard Lyon1, 69622 Villeurbanne Cedex, France  
Phone: +33 4 72 44 48 83  
Fax: +33 4 72 43 13 12

## Abstract

This paper proposes a novel adaptive algorithm to extract facial feature points automatically such as eyebrows corners, eyes corners, nostrils, nose tip, and mouth corners in frontal view faces, which is based on cumulative histogram approach by varying different threshold values. At first, the method adopts the Viola-Jones face detector to detect the location of face and also crops the face region in an image. From the concept of the human face structure, the six relevant regions such as right eyebrow, left eyebrow, right eye, left eye, nose, and mouth areas are cropped in a face image. Then the histogram of each cropped relevant region is computed and its cumulative histogram value is employed by varying different threshold values to create a new filtering image in an adaptive way. The connected component of interested area for each relevant filtering image is indicated of respective feature region. A simple linear search algorithm for eyebrows, eyes and mouth filtering images and contour algorithm for nose filtering image are applied to extract our desired corner points automatically. The method was tested on a large BioID frontal face database in different illuminations, expressions and lighting conditions and the experimental results have achieved average success rates of 95.27%.

**Keywords:** *Connected Component, Corner Point Detection, Face Recognition, Cumulative Histogram, Linear Search.*

## 1. Introduction

Face analysis such as facial features extraction and face recognition is one of the most flourishing areas in computer vision like identification, authentication, security, surveillance system, human-computer interaction,

psychology and so on [1]. Facial features extraction is the initial stage for face recognition in the field of vision technology. The most significant feature points are eyebrows corners, eyes corners, nostrils, nose tip, and mouth corners. These are the key components for face recognition [2], [3]. Eyes are the most crucial facial feature for face analysis because of its inter-ocular distance, which is constant among people and unaffected by moustache or beard [3]. Eyebrows, eyes and mouth also convey facial expressions. Another valuable face feature points are nostrils because nose tip is the symmetry point of both right and left side face regions and nose indicates the head pose and it is not impacted by facial expressions [4]. Therefore, face recognition is distinctly influenced by these feature points.

Currently, Active Shape Model (ASM) and Active Appearance Model (AAM) are extensively used for face alignment and tracking [5]. Facial feature extraction methods could be divided in two categories: texture-based and shape-based methods.

Texture-based methods take local texture e.g. pixel values around a given specific feature point instead of concerning all facial feature points as a shape (shape-based methods). Some texture-based facial feature extraction algorithms are: hierarchical 2-level wavelet networks for facial feature localization [6], facial point detection using log Gabor wavelet networks by employing geometry cross-ratios relationships[7], neural-network-based eye-feature detector by locating micro-features instead of entire eyes[8]. Some shape-based facial feature extraction

algorithms including AAM, based on face detectors are:

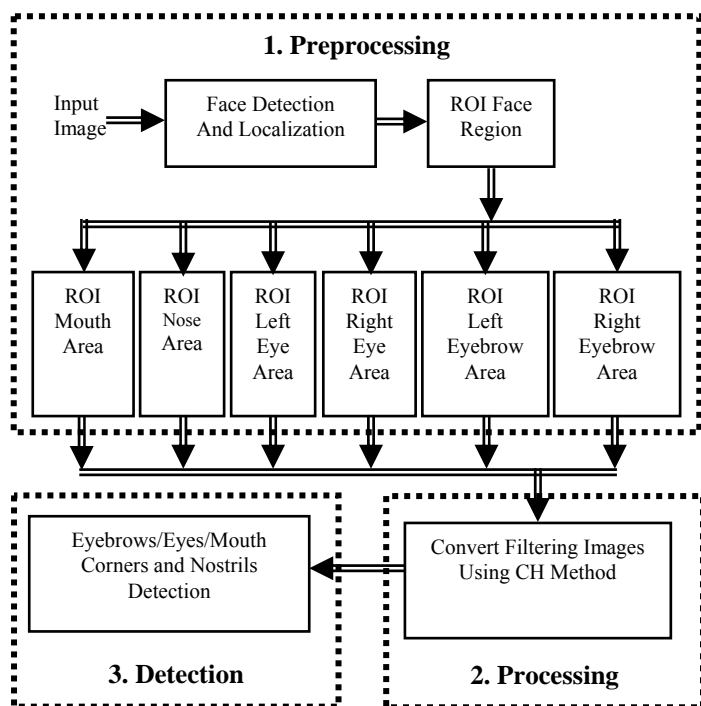


Figure 1. Block diagram of proposed feature extraction algorithm.

view-based active wavelet network [9], view-based direct appearance models [10]. The combination of texture- and shape-based algorithms are: elastic bunch graph matching [11], AdaBoost with Shape Constrains [12], 3D Shape Constraint using Probabilistic-like Output [13]. Wiskott et al. [11] represented faces by a rectangular graph which is based on Gabor wavelet transform and each node labelled with a set of complex Gabor wavelet coefficients. Cristinacce and Cootes [12] used the Haar features based AdaBoost classifier combined with the statistical shape model. In both ASM and AAM, a model is built for predefined points by using the test images and then an iterative scheme is applied to this model in detecting feature points. Most of the above mentioned algorithms are not entirely reliable due to variation in pose, illumination, facial expression, and lighting condition and high computational complexity. So, it is indispensable to develop robust, automatic, and accurate facial feature point localization algorithms, which are capable in coping different imaging conditions.

In this paper, we propose a robust adaptive algorithm based on cumulative histogram (CH) scheme that extracts the facial feature points in a fast as well as accurate way under varying illuminations, expressions and lighting conditions. Figure 1 shows the block diagram of our

proposed algorithm that includes preprocessing, main processing and detection blocks. The preprocessing block detects the face and crops the face, right eyebrow, left eyebrow, right eye, left eye, nose, and mouth areas. The processing block is responsible for six ROIs such as right eyebrow, left eyebrow, right eye, left eye, nose, and mouth areas and then converts into binary images. The detection block detects the corner points of six ROIs. The remainder of the paper is organized as follows. Section 2 describes the region of interest (ROI) detection. In section 3, we present the mathematical description of CH method, which form the basis for our approach, and then we explain the facial feature point detection with the algorithmic details. Section 4 shows the experimental results of our facial feature extraction system. Finally we conclude the paper along with highlighting future work directions in section 5.

## 2 Region of Interest Detection

A rectangular portion of an image to perform some other operation and also to reduce the computational cost for further processing is known as region of interest (ROI).

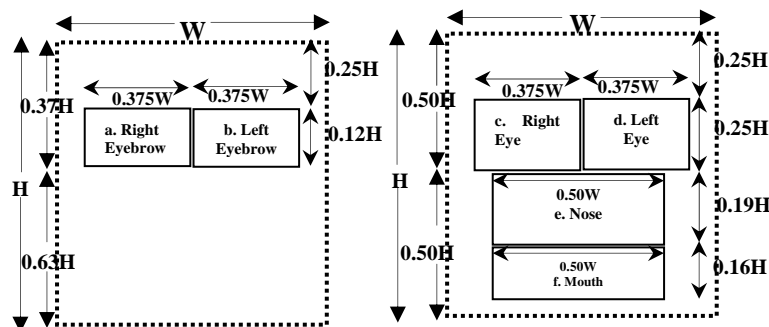


Figure 2. Location and size of six ROIs of a face image such as (a.) Right Eyebrow (Size:  $0.375W \times 0.12H$ ), (b.) Left Eyebrow (Size:  $0.375W \times 0.12H$ ), (c.) Right Eye (Size:  $0.375W \times 0.25H$ ), (d.) Left Eye (Size:  $0.375W \times 0.25H$ ), (e.) Nose (Size:  $0.50W \times 0.19H$ ), and (f.) Mouth (Size:  $0.50W \times 0.16H$ ) where,  $W$ =Image Width and  $H$ =Image Height.

By applying the Viola-Jones face detector algorithm, the detected face region is cropped first then we divide the face area vertically into upper, middle and lower parts [14]. From the human frontal face structure concept, eyebrows & eyes, nose, and mouth areas are situated in upper, middle, and lower portions of the face image, respectively. Again, the upper portion is partitioned horizontally into left and right segments for isolating right-eyebrow and right-eye and also left-eyebrow and left-eye, respectively.

Finally, the smallest ROI regions are segmented for right-eyebrow, right-eye, left-eyebrow, left-eye, nose, and

mouth in order to increase the detection rate. Figure 1, Figure 2, and Figure 3(d) are shown the block diagram of our proposed algorithm, location and size of six ROIs and cropped images, respectively.

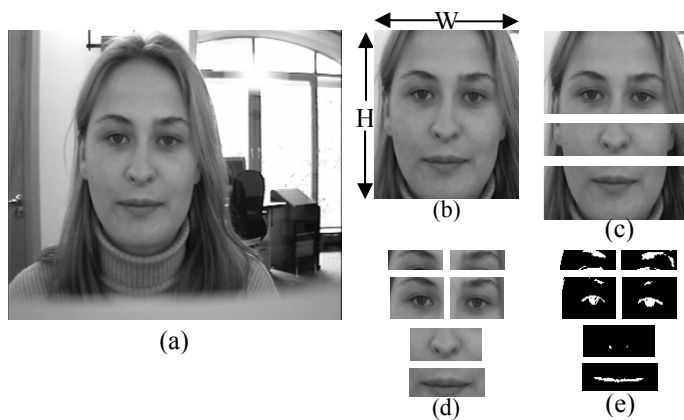


Figure 3. Procedure of our proposed algorithm: (a) Input image, (b) Detected and cropped the face, (c) Face is divided into three vertical parts, which are indicated eyebrows, eyes, nose and mouth areas, (d) Six ROIs show the exact right-eyebrow, right-eye, left-eyebrow, left-eye, nose and mouth regions, (e) Applying CH method, all of the six ROIs are converted into new filtering images.

### 3. Facial Features Extraction

Our proposed method exhibits the location of twelve crucial feature points including eight corner points for right-eyebrow, right-eye, left-eyebrow, left-eye, two points for nostrils and two corner points for mouth as shown in the Figure 3(d) and Figure 3(e). Feature points are extracted by an adaptive approach. To create the new filtering (binary) images, the following mathematical concepts are applied on each of the six original cropped (ROIs) gray scale images such as right-eyebrow, right-eye, left-eyebrow, left-eye, nose and mouth regions(see Figure 3(d) and Figure 3(e))[16],[17].

$$P_{I(x,y)}(v) = P(I(x,y)=v) = \frac{n_v}{N} \quad \text{Where, } 0 \leq v \leq 255 \quad (1)$$

$$CH_{I(x,y)}(v) = \sum_{i=0}^v P_{I(x,y)}(i) \quad (2)$$

$$I_{FI}(x,y) = \begin{cases} 255 & \text{when } CH(I(x,y)) \leq Th \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Where,  $I(x,y)$  is denoted by each of the six original cropped gray scale images,  $P_{I(x,y)}(v)$  is the histogram representing probability of an occurrence of a pixel of gray level  $v$ ,  $n_v$  is the number of pixels having each pixel

value is  $v$  and  $N(\text{width} \times \text{height})$  is the total number of pixels, and  $CH_{I(x,y)}(v)$  is the cumulative histogram(CH) function up to the gray level  $v$  for an image  $I(x,y)$ [16],[17], where  $0 \leq v \leq 255$ . The CH ( $v$ ) is measured by summing up the all histogram values from gray level 0 to  $v$  [20]. The new filtering image,  $I_{FI}(x,y)$  is achieved when CH value is not exceeded the threshold value  $Th$  and the  $I_{FI}(x,y)$  image only contains the white pixels of our specific desired connected component area. Figure 3(e) is shown the respective white pixel's connected component of all filtering images for right-eyebrow, right-eye, left-eyebrow, left-eye, nose, and mouth region. Three different groups of threshold values are used for our evaluation purpose. One for eyebrows region ( $0.01 \leq Th \leq 0.25$ ) another for eyes and mouth regions ( $0.01 \leq Th \leq 0.10$ ) and the other for nose region ( $0.001 \leq Th \leq 0.010$ ) because nostrils contain minimum numbers of low intensity pixels of original image compare to eyebrows, eyes and mouth region (see Figure 5) [4].

#### 3.1 Eyebrow, Eye and Mouth Corner Points Detection

A simple linear search concept is applied on right-eyebrow, right-eye, left-eyebrow left-eye, and mouth filtering images to detect the first white pixel locations as the candidate points:

- (1) Starting from top-left position for right corner points and (2) starting from top-right position for left corner points to search downward direction for eyebrows corner points
- (3) Starting from bottom-left position for right corner points and (4) starting from bottom-right position for left corner points to search upward direction for eyes and mouth corner points.

The located first white pixel's positions are the candidate corner points.

#### 3.2 Nostrils Detection and Nose Tip Calculation

A contour algorithm, using connected component, is applied on nose filtering image to select the last (right nostril) and the previous last (left nostril) contours from bottom to upward direction. Then the last and the previous last contour's element locations are sorted as an ascending order according to horizontal direction(x-value). The locations of the last element (right nostril point) of the last contour and the first element (left nostril point) of the previous last contour are the candidate nostrils. Nose tip is computed as the mid point between nostrils because the nose tip conveys the highest gray scale value so that nose filtering image shows insufficient information about it(see

the 2nd filtering image started from the lower position of Figure 3(e))[6], [18].

All of the detected twelve corner points are indicated as 'black plus symbols', and only calculated nose tip is indicated as 'black solid circle' as shown in Figure 6.

### 3.3 Proposed Algorithm

The proposed algorithm is organized by three sections, which are included "preprocessing", "processing", and "detection" sections (see Figure 1). The preprocessing section detects the face and its location and then crops the face, right-eyebrow, right-eye, left-eyebrow, left-eye, nose, and mouth regions in an image. We assume that as a frontal face image, the eyebrows & eyes, nose, and mouth are located in upper half, middle and lower parts, respectively, in an image (see Figure 2 and Figure 3). In the processing section, the cropped images i.e. six ROIs such as right-eyebrow, right-eye, left-eyebrow, left-eye, nose, and mouth are converted into filtering images by applying CH method (using equations (1),(2), & (3)) [16],[17]. Applying simple linear search and contour concepts on these filtering images, the detection section finds out the all facial feature points such as right & left eyebrows corners, right & left eyes corners, nostrils, and mouth corners. The step by step procedures of our proposed algorithm are described as follows.

#### A. Preprocessing section

1. Input:  $I_{\text{whole-face-window}}(x,y)$  =Frontal face gray scale image having head and shoulder (whole face window)(see Figure 3(a)).
2. Detect and localize the face by applying the OpenCV face detection algorithm [19].
3. Detect the regions of interest (ROI) for face, right-eyebrow, right-eye, left-eyebrow, left-eye, nose, and mouth by applying the OpenCV ROI library functions [19] and then we build the following new images.

(a)  $I_{\text{face}}(x,y)$  =New image having only face area and its size is  $W \times H$ (see Figure 2 and Figure 3(b)) Where,  $W$ =image width,  $H$ =image height.

(b)  $I_{\text{eyebrow-right}}(x,y)$  =New image having only right eyebrow area and its size is  $0.375W \times 0.12H$ (See Figure 2 and Figure 3(d)).

(c)  $I_{\text{eyebrow-left}}(x,y)$  =New image having only left eyebrow area and its size is  $0.375W \times 0.12H$ (See Figure 2 and Figure 3(d)).

(d)  $I_{\text{eye-right}}(x,y)$  =New image having only right eye area and its size is  $0.375W \times 0.25H$ (See Figure 2 and Figure 3(d)).

(e)  $I_{\text{eye-left}}(x,y)$  =New image having only left eye area and its size is  $0.375W \times 0.25H$ (See Figure 2 and Figure 3(d)).

(f)  $I_{\text{nose}}(x,y)$  =New image having only nose area and its size is  $0.50W \times 0.19H$ (See Figure 2 and Figure 3(d)).

(g)  $I_{\text{mouth}}(x,y)$  =New image having only mouth area and its size is  $0.50W \times 0.16H$ (See Figure 2 and Figure 3(d)).

#### B. Processing section

4. Apply CH method(using equations (1),(2), & (3)) [16],[17] on the above six ROIs such as  $I_{\text{eyebrow-right}}(x,y)$ ,  $I_{\text{eyebrow-left}}(x,y)$ ,  $I_{\text{eye-right}}(x,y)$ ,  $I_{\text{eye-left}}(x,y)$ ,  $I_{\text{nose}}(x,y)$ , and  $I_{\text{mouth}}(x,y)$  images(see Figure 3(d)) and convert it into new filtering(binary) images such as  $I_{\text{FI\_eyebrow-right}}(x,y)$ ,  $I_{\text{FI\_eyebrow-left}}(x,y)$ ,  $I_{\text{FI\_eye-right}}(x,y)$ ,  $I_{\text{FI\_eye-left}}(x,y)$ ,  $I_{\text{FI\_nose}}(x,y)$ , and  $I_{\text{FI\_mouth}}(x,y)$  for different threshold values(see Figure 3(e)).

#### C. Detection section

- 5.(a)A simple linear search concept is applied on filtering images such as  $I_{\text{FI\_eyebrow-right}}(x,y)$ ,  $I_{\text{FI\_eyebrow-left}}(x,y)$ ,  $I_{\text{FI\_eye-right}}(x,y)$ ,  $I_{\text{FI\_eye-left}}(x,y)$ , and  $I_{\text{FI\_mouth}}(x,y)$  for eyebrows, eyes and mouth corner points, then find out the first white pixel location as top-down approach on eyebrows filtering images and bottom-up approach on eyes & mouth filtering images. To locate for all corner points :

(1) Starting searches from top-left position for right corner points and (2) starting searches from top-right position for left corner points to search top-down approach for eyebrows corner points.

(3) Starting searches from bottom-left position for right corner points and (4) starting searches from bottom-right position for left corner points to search bottom-up approach for eyes & mouth corner points.

(b) Apply the OpenCV contour library function on filtering image,  $I_{\text{FI\_nose}}(x,y)$  for nostrils; then consider the locations of the last element(right nostril point) and the first element (left nostril point) for the last and the previous last contours as a bottom-up approach where, the contour's element locations are sorted horizontally (x-value) as an ascending order[19]. The sorted locations of contour elements for the last (Contour P) and the previous last (Contour Q) contours are shown in figure 4.



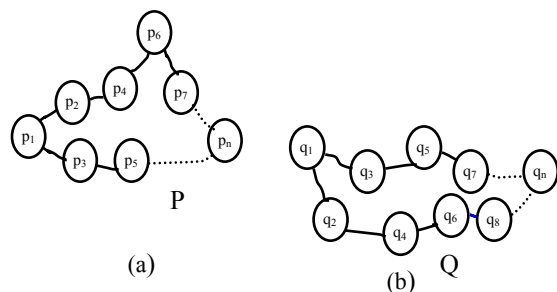


Figure 4. The locations of contours elements of Nose filtering image : (a) Sorted locations(x-value) of the last contour elements= $P(p_1, p_2, \dots, p_n)$ , (b) Sorted locations(x-value) of the previous last contour elements= $Q(q_1, q_2, \dots, q_n)$ .

**The locations of the last contour elements are:**

$p_1(x_{11}, y_{11}), p_2(x_{12}, y_{12}), \dots, p_n(x_{1n}, y_{1n})$ . Where,  $x_{21} < x_{12} < \dots < x_{1n}$ .

**Right Nostril**=The last element of the last contour =  $p_n(x_{1n}, y_{1n})$

**The locations of the previous last contour elements are:**

$q_1(x_{21}, y_{21}), q_2(x_{22}, y_{22}), \dots, q_n(x_{2n}, y_{2n})$ . Where,  $x_{21} < x_{22} < \dots < x_{2n}$ .

**Left Nostril**=The first element of the previous last contour =  $q_1(x_{21}, y_{21})$

(c) The mid point (x-value) between nostrils and minimum y-value for nose tip is:

$$X_{nose\ tip} = \text{Integer value of } (x_{1n} + x_{21}) / 2$$

$$Y_{nose\ tip} = ((y_{1n} < y_{21}) ? y_{1n} : y_{21}) - 8$$

**So Nose Tip** =  $(X_{nose\ tip}, Y_{nose\ tip})$

At last, the detected points are transferred to the  $I_{face}(x,y)$  image (see Figure 3(b) and Figure 6).

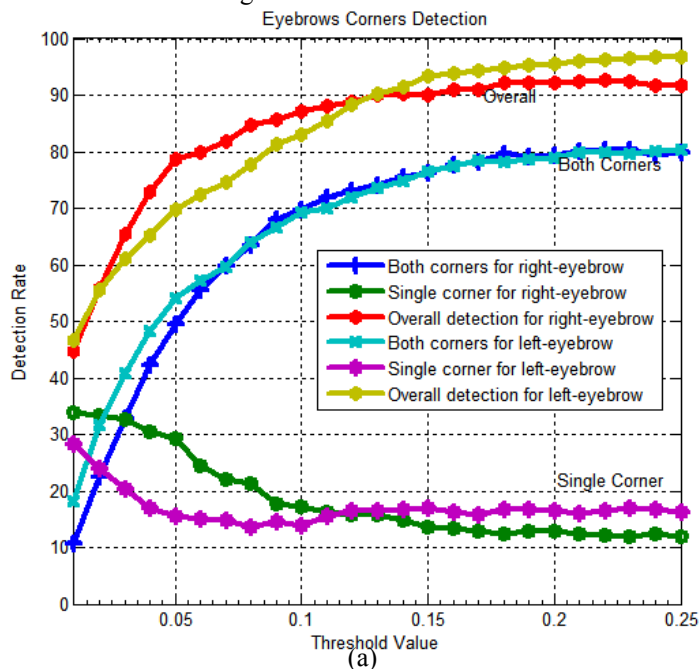
## 4 Experimental Results

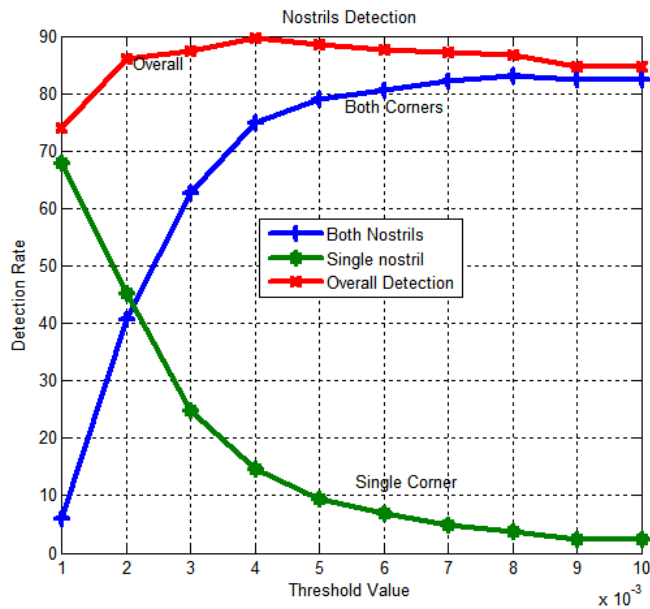
### 4.1 Face Database

The work described in this paper is used head-shoulder BioID face database [15]. The dataset with the frontal view of a face of one out of 23 different test persons consists of 1521 gray level images having properties of different illumination, face area, complex background with a resolution of 384x286 pixel. During evaluation, some images are omitted due to : (1) detecting false region (not face) by Viola-Jones face detector [14] and (2) person with large size eye glasses and highly dense moustache or beard as a complex background property of an image.

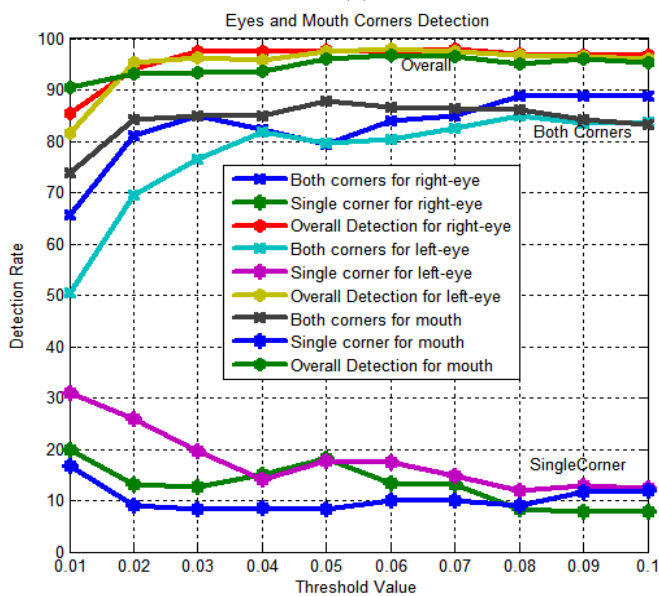
### 4.2 Results

The proposed algorithm was primarily developed and tested on Code::Blocks the open source, cross-platform combine with c++ language, and GNU GCC compiler. Some OpenCV library functions were used for face detection and localization, cropping and also connected component (contour algorithm) purpose [19]. During evaluation, three different groups of threshold values were used for our CH analysis (using equations (1), (2), & (3)) [16], [17]. One is  $0.01 \leq Th \leq 0.25$  for locating eyebrows corner points, another is  $0.01 \leq Th \leq 0.10$  for locating eyes and mouth corner points and the other is  $0.001 \leq Th \leq 0.010$  for locating nostrils. Figure 5 shows the detection rate of twelve corner points by using different threshold values. Figure 5(a) shows single corner, both corners and overall detection rate for right eyebrow, left eyebrow corner points, Figure 5(b) shows single nostril, both nostrils and overall detection rate for nostrils and Figure 5(c) shows single corner, both corners and overall detection rate for right eye, left eye, and mouth corner points. The combination of single corner and both corners detection rate is considered as the overall detection rate. Threshold values 0.220, 0.240, 0.070, 0.060, 0.004, and 0.060 produce the detection rates 92.56%, 96.83% 97.92%, 98.02%, 89.58%, and 96.73% for right-eyebrow corners, left-eyebrow corners, right-eye corners, left-eye corners, nostrils, and mouth corners, respectively. Table 1 indicates the results of our facial feature extraction algorithm, where the overall an average detection rate is 95.27%. We compared our algorithm with R.S. Feris, et al. [6] and D. Vukadinovic, M. Pantic [2]. The comparison results are shown in table 2. Some of the detection results are shown in the Figure 6.





(b)



(c)

Figure 5. Detection Rate using different threshold values of CH method on BioID face database: (a) Eyebrows Detection Curves (Single, Both, Overall), (b) Nostrils Detection Curves (Single, Both, Overall), (c) Eyes and Mouth Corners Detection Curves (Single, Both, Overall).

Using the six relevant regions of a frontal view face image such as right eyebrow, left eyebrow, right eye, left eye, nose, and mouth areas are shown an average detection rate is 95.27% (See the table 1), whereas using the four relevant regions such as right eye, left eye, nose, and mouth areas are shown an average detection rate is 95.56% [15], [20].

Table 1: Table of feature points detection rate

Table 2: Comparisons with 2-level GWN [6] and GFBBC [2]

Features	Detection Rate (%) for both Points/Corners	Detection Rate (%) for single Point/Corner	Overall Detection Rate (%)	Threshold Value for CH Method
Right Eyebrow	80.36	12.20	92.56	0.220
Left Eyebrow	80.16	16.67	96.83	0.240
Right Eye	84.82	13.10	97.92	0.070
Left Eye	80.46	17.56	98.02	0.060
Nostrils	75.00	14.58	89.58	0.004
Mouth Corners	86.71	10.02	96.73	0.060
<b>Average</b>	<b>81.25</b>	<b>14.02</b>	<b>95.27</b>	-

Algorithms	Average Detection rate (%)
2-level GWN	92.87
GFBBC	93.00
Ours	95.27



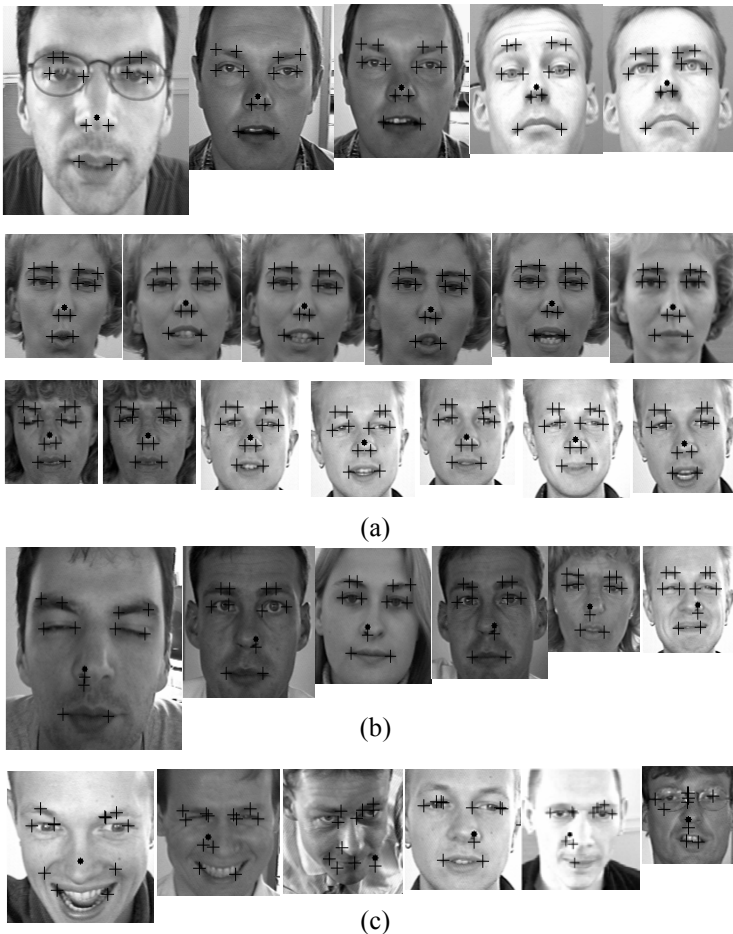


Figure 6. Result of detected feature points:(a) Some true detection, (b) Some single nostril detection and (c) Some false detection.

## 5 Conclusions and Future Work

In this paper, we have shown how salient facial features are extracted based on cumulative histogram CH method in an adaptive manner combined with face detector, simple linear search, and also connected component concepts i.e. contour algorithm in various expression and illumination conditions in an image. Image segments are converted into filtering images with the help of CH approach by varying different threshold values instead of applying morphological operations. Our algorithm was assessed on free accessible BioID gray scale frontal face database. The experimental results confirmed the higher detection rate as compare to other well known facial feature extraction algorithms.

Future work will concentrate to improve the detection rate of both corner points instead of single corner point by

using a single threshold group instead of multiple threshold groups and face recognition, as well.

## References

- [1] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face Recognition: A Literature Survey", *ACM Computing Surveys*, Vol. 35, No. 4, December 2003.
- [2] D. Vukadinovic and M. Pantic, "Fully Automatic Facial Feature Point Detection Using Gabor Feature Based Boosted Classifiers", *2005 IEEE International Conference on Systems, Man and Cybernetics Waikoloa, Hawaii*, October 10-12, 2005.
- [3] <http://eprints.um.edu.my/877/1/GS10-4.pdf>
- [4] Wei Jen Chew, Kah Phooi Seng, and Li-Minn Ang, "Nose Tip Detection on a Three-Dimensional Face Range Image Invariant to Head Pose", *Proceedings of The International Multi Conference of Engineers and Computer Scientists 2009*, Vol I, IMECS 2009, March 18-20, 2009, Hong Kong.
- [5] I. Matthews and S. Baker, "Active Appearance Models Revisited", *Int'l Journal Computer Vision*, vol. 60, no. 2, pp.135-164, 2004.
- [6] R.S. Feris, et al., "Hierarchical Wavelet Networks for Facial Feature Localization", *Proc. IEEE Int'l Conf. Face and Gesture Recognition*, pp. 118-123, 2002.
- [7] E. Holden, R. Owens, "Automatic Facial Point Detection", *Proc. The 5<sup>th</sup> Asian Conf. on Computer Vision, 23-25 January 2002, Melbourne, Australia*.
- [8] M. J. T. Reinders, et al., "Locating Facial Features in Image Sequences using Neural Networks", *Proc. IEEE Int'l Conf. Face and Gesture Recognition*, pp.230-235, 1996.
- [9] C. Hu, et al., "Real-time view-based face alignment using active wavelet networks", *Proc. IEEE Int'l Workshop Analysis and Modeling of Faces and Gestures*, pp. 215-221, 2003.
- [10] S. Yan, et al., "Face Alignment using View-Based Direct Appearance Models", *Int'l J. Imaging Systems and Technology*, vol. 13, no. 1, pp. 106-112, 2003.
- [11] L. Wiskott, et al., "Face Recognition by Elastic Bunch Graph. Matching" *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no.7, pp. 775-779, 1999.
- [12] D. Cristinacce, T. Cootes, "Facial Feature Detection Using AdaBoost with Shape Constrains", *British Machine Vision Conference*, 2003.
- [13] L. Chen, et al., "3D Shape Constraint for Facial Feature Localization using Probabilistic-like Output", *Proc. IEEE Int'l Workshop Analysis and Modeling of Faces and Gestures*, pp. 302-307, 2004.
- [14] P. Viola and M. J. Jones, "Robust Real-time Object Detection", *International Journal of Computer Vision*, Vol. 57, No.2, p.137-154, 2004.
- [15] BioID Face Database, Available: <http://www.bioid.com/downloads/facedb/index.php>
- [16] Joung-Youn Kim, Lee-Sup Kim, and Seung-Ho Hwang, "An Advanced Contrast Enhancement Using Partially Overlapped Sub-Block Histogram", *IEEE Transactions On*

*Circuits And Systems For Video Technology*, Vol. 11, No. 4, April 2001.

- [17] Mansour Asadifard, and Jamshid Shanbezadeh, "Automatic Adaptive Center of Pupil Detection Using Face Detection and CDF Analysis", *Proceedings of The International Multi Conference of Engineers and Computer Scientists 2010* , pp130-133 , IMECS 2010,17-19 March, 2010, Hong Kong.
- [18] S. Jahanbin, et al., "Automated Facial Feature Detection from Portrait and Range Images", *SSIAI '08 Proc. Of the 2008 IEEE Southwest Symposium on Image Analysis and Interpretation*, 24-26 March 2008.
- [19] <http://sourceforge.net/projects/opencvlibrary/files/opencv-win/2.0/OpenCV-2.0.0a-win32.exe/download>
- [20] Sushil Kumar Paul, Saida Bouakaz and Mohammad Shorif Uddin, "Automatic Adaptive Facial Feature Extraction using CDF Analysis", *Proceedings of The International Conference on Digital Information and Communications Technology and its Applications, Communications in Computer and Information Science*, Volume 166, Part 1, Pages 327-338, DICTAP2011, 21-23 June 2011, Dijon, France.



**Sushil Kumar Paul** is currently pursuing his PhD in Face Recognition, Department of Computer Science and Engineering (CSE) from Jahangirnagar University (JU), Dhaka. He has completed his Master of Science in CSE from United International University (UIU) and Bachelor of Science in Electrical and Electronic Engineering from Bangladesh University of Engineering and Technology (BUET), Dhaka. He has

worked as an eLINK(East-West Link for Innovation, Networking and Knowledge Exchange) scholar in LIRIS Laboratory, University Claude Bernard Lyon1, France in 2010. He has twelve years of teaching experience as an instructor (Computer/Electronics) in many Polytechnic Institutes under Directorate of Technical Education (DTE) in Bangladesh. He has also been participated many short terms foreign & local training programs and he is currently working as Principal, Narsingdi Polytechnic Institute under DTE. He has published one international conference paper and also worked as a reviewer. He is a writer of one local textbook and a head examiner under Bangladesh Technical Education Board (BTEB). His research interests are in computer vision, digital imaging, machine learning, and data mining including face, facial expression and motion recognition. He is also a member of IEB, BCS and BUET87 foundation.



**Mohammad Shorif Uddin** received his PhD in Information Science from Kyoto Institute of Technology, Japan, Masters of Education in Technology Education from Shiga University, Japan and Bachelor of Science in Electrical and Electronic Engineering from Bangladesh University of Engineering and Technology (BUET). He joined the Department of Computer Science and Engineering, Jahangirnagar University, Dhaka in

1992 and currently serves as a Professor of this department. He began his teaching career in 1991 as a Lecturer of the Department

of Electrical and Electronic Engineering, Chittagong University of Engineering and Technology (CUET). He undertook postdoctoral research at Bioinformatics Institute, A-STAR, Singapore, Toyota Technological Institute, Japan and Kyoto Institute of Technology, Japan. His research is motivated by applications in the fields of computer vision, pattern recognition, blind navigation, bio-imaging, medical diagnosis and disaster prevention. He has published a remarkable number of papers in peer-reviewed international journals and conference proceedings. He holds two patents for his scientific inventions. He received the Best Presenter Award from the International Conference on Computer Vision and Graphics (ICCVG 2004), Warsaw, Poland. He is the co-author of two books. He is also a member of IEEE, SPIE, IEB and a senior member of IACSIT.

**Saida Bouakaz** received her PhD from Joseph Fourier University in Grenoble, France. She is currently a professor in the Computer Science and head of the SAARA Research Team, LIRIS Lab at Claude Bernard University, Lyon1. Her research interests are in computer vision and graphics including motion capture and analysis, face, facial expression and gesture recognition, facial animation.



# Validity Index and number of clusters

Mohamed Fadhel SAAD<sup>1</sup> and Adel M. ALIMI<sup>1</sup>

<sup>1</sup> Research Group on Intelligent Machines, University of Sfax,  
ENIS Sfax, 3038, Sfax, Tunisia

## Abstract

Clustering (or cluster analysis) has been used widely in pattern recognition, image processing, and data analysis. It aims to organize a collection of data items into  $c$  clusters, such that items within a cluster are more similar to each other than they are items in the other clusters. The number of clusters  $c$  is the most important parameter, in the sense that the remaining parameters have less influence on the resulting partition. To determine the best number of classes several methods were made, and are called validity index. This paper presents a new validity index for fuzzy clustering called a Modified Partition Coefficient And Exponential Separation (MPCAES) index. The efficiency of the proposed MPCAES index is compared with several popular validity indexes. More information about these indexes is acquired in series of numerical comparisons and also real data Iris.

**Keywords:** Fuzzy clustering, Fuzzy c-means, Validity index.

## 1. Introduction

Fuzzy classification algorithms require the user to predefine the number of clusters ( $c$ ), but it is not always possible to know this number in advance. Since the scores obtained using the  $c$ -means family algorithms depend on the choice of  $c$ , it is necessary to validate each result of the partitions once they are found. This validation is performed by a specific algorithm that allows to assume the appropriate value of the number  $c$ . We call this algorithm "validity index of the classification". It evaluates each class and determines the optimal or valid partition.

During the last years, it has been proposed many validity indexes. Most of them came from different studies on the number of classes. Among these indexes, there are two important types for  $c$ -means: one is based on the fuzzy partition of the dataset and the other is based on the geometric structure.

The main idea of the validity functions based on fuzzy partitioning: less fuzziness partitioning is more the performance is better. The representative functions for these are the coefficient of partitioning  $V_{pc}$  (Validity partition coefficient) [1] and the entropy of partitions  $V_{pe}$  (Validity partition entropy) [2]. Empirical studies [3] think that the maximum  $V_{pc}$  and minimum  $V_{pe}$  lead to a correct interpretation of the samples considered. The best

performance is achieved when the  $V_{pc}$  gets its maximum

value or  $V_{pe}$  obtains its minimum.

Note that in some cases these functions validity cannot obtain their optimal values simultaneously. In the next sections, we detail the algorithms of the most recent validity index functions.

## 2. Presentation of validity index

Classification Validity Indexes (CVIs) have attracted the attention of researchers in order to validate the partition found by the  $c$ -means algorithm. The CVIs can signal the perfect input parameters with the best results by taking a minimum (or maximum). The quality of the result is incorporated in the number of classes and purity of each class. Purity is the sum of data objects in the majority class and this for each partition found. The number of classes is related to the purity of these classes. Thus, if the number of classes  $c$  is right, there is a high purity.

Several conventional CVIs have been developed with new instance types of intra-class and inter-class. However, the fundamentals for designing the CVIs were rarely defined in a clear manner.

### 2.1 Background

Historically, the classification validity indexes related to the  $c$ -means family algorithms have been proposed, first is the partitioning coefficient  $V_{pc}$  and entropy scores  $V_{pe}$  developed by Bezdek, as described in previous section.

The disadvantages of the coefficients  $V_{pc}$  and  $V_{pe}$  are the lack of direct connection to the geometrical structure of data, and their tendency to decrease with the number  $c$ . Moreover, the main idea of the functions of validity is based on the geometry of objects, within the same class must be compact and in different classes should be separated. The coefficient of separation proposed by Gunderson in 1978 [4] was the first validity index that reflects explicitly the geometric properties of data.

Another remedy these drawbacks have been made in the function of Fukuyama and Sugeno [5], density classes.

proposed by Gath [6] and the function of Xie and Beni [7]. It is expected that the reduction of these functions at least, leads to good classification. Intuitively, the lack of clarity and compactness of a classification should decrease with increasing number of classes. For example, the partition entropy decreases to zero when  $c$  becomes very large and tends to the number of objects  $n$ . For this reason, the validity indices take as the maximum number of classes, the square root of the number of items:  $c_{max} = \sqrt{n}$ . Once the partition is obtained by exact or fuzzy classification methods, the validity index can help determine the reliability of this partition for the data structure.

There may be mentioned the best-known index:

- Partition coefficient

$$V_{pc} = \frac{\sum_{i=1}^c \sum_{j=1}^n \mu_{ij}^2}{n} \quad (1)$$

- Partition entropy

$$V_{pe} = \frac{\sum_{i=1}^c \sum_{j=1}^n \mu_{ij} \log_2(\mu_{ij})}{n} \quad (2)$$

- Fukuyama Sugeno

$$V_{fs} = \sum_{i=1}^c \sum_{j=1}^n \mu_{ij} (d^2(x_{pj}, v_i) - d^2(x_{pj}, v)) \quad (3)$$

- Partition coefficient

$$V_{pb} = \frac{\sum_{i=1}^c \sum_{j=1}^n \mu_{ij}^2 d^2(x_{pj}, v_i)}{n(\min_{i \neq k} d^2(v_i, v_k))} \quad (4)$$

To find the optimal partition, we must maximize  $V_{pc}$  or minimize  $V_{pe}$ ,  $V_{fs}$ ,  $V_{pb}$ .

Indices are classified into two types: addition and report type. The type is determined by how the intra-class and inter-class distance are coupled. Depending on the combination of these two distances, the results of indices validating classification carried out, are distinguished in connection with the domain structure of data having different aspects.

For several CVIs, the mean is a step of calculating intra-class distances. The average implies input values and gives a summary value on compactness. Therefore, this may mask the discriminatory ability of CVIs. Thus, the formulas introduced in the indices must use techniques that apply to all areas. We define for this aspect of fields of data such as compactness, separability, noise and overlap.

- **The compactness:** is a measure of the proximity of points vectors comprising the same class of its center.

- **The separability:** indicates how two classes are distinct and isolated from one of other. The separation gives the distance between two different classes.

Most validation indices proposed in recent years, including the index of Xie and Beni [7] and Davies-Bouldin index [8], are

focused on two properties: compactness and separation. Thus, a smaller local value shows that each class is compact and great value in separation of the classes well separated.

- **The noise:** a noisy environment has points parasites that do not belong to any class of dataset. Most validity indices measure the degree of compactness and separation for the dataset and then find an optimal number of classes. If the dataset contains some noise points, then we can see that the validity indices take the noisy point in a compact and separated class from the rest of the classes. Thus, the noise aspect is crucial in the classification of data.

- **The overlap:** a measure indicating the degree to which two classes overlap and have similar feature vectors in common. This is defined between the fuzzy classes by calculating an overlap of inter-class. A better score is obtained at a minimal degree of overlap.

If a noise point is considered, a parasitic class well-identified in dataset. The found partition does not properly describe the data structure. Thus, the noise points that exist in different environments should not have enough good opportunities to be valid classes. The compactness is a measure of variation or dispersion of data in one class, and separation is an indicator of the isolation of classes from each other. A conventional approaches measuring compactness cannot clearly distinguish the different classes composing the dataset. In fact, compactness is a distance factor vector points in the center and degrees of membership of these items to class.

If the distance  $\|x_{pj} - x_{pk}\|$  is great, the membership degree of point  $x_{pj}$  to the center  $v_j$  is small (case of the first class (a) in Figure 1). Else, if the point  $x_{pj}$  is near the center  $v_j$ , the distance between them is small and the membership degree is important (case of the first class (b) in Figure Fig. 1). Thus, the compactness values of the two classes are similar and do not reflect the geometry of the dataset. In addition, conventional measures of separation have limited ability to differentiate between geometric structures of the classes, because the calculation is based solely on information center and does not consider the overall shape of the classes as is shown schematically in Figure Fig. 2. In fact, there are two identical values of separation for pair of classes with different forms.

## 2.2 Validity indices based on the separation

This index is developed by Tsekoura and Sarimveis in 2004 [9] to validate the FCM algorithm. Its function takes a compactness measure to describe the change of classes, and introduces the concept of fuzzy separation to determine the isolation of groups. The basic design of separation is the deviation between two fuzzy cluster centers. This index called Separation Validity Index (SVI) is based on compactness and separation criteria. A total compactness quantity is used to describe similarities

between multiple objects in the same class, and a separation measure provides an evaluation of distances between the cluster centers when they are calculated relative to each other.

The Formula of SVI validity index is as follows:

$$SVI = \frac{F}{S} \quad (5)$$

The overall compactness of classification  $F$  is the sum of all the compactness  $F_i$ , where  $i$  is the class index ( $1 \leq i \leq c$ ).

The compactness of the class  $c_i$  is given by:

$$F_i = \frac{\sigma_i}{n_i}, \quad 1 \leq i \leq c \quad (6)$$

The variance and the cardinality of the class  $c_i$  are given respectively by:

$$\sigma_i = \sum_{j=1}^{n_i} \mu_{ij}^2 d^2(x_j, v_i) \quad (7)$$

$$n_i = \sum_{j=1}^{n_i} \mu_{ij}^2, \quad 1 \leq i \leq c \quad (8)$$

With  $\mu_{ij}$  is the membership degree of vector  $x_j$  to the cluster  $c_i$  and  $v_i$  is the center of cluster  $c_i$ .

The overall separation of  $c$  classes is given by the equation:

$$S = \sum_{i=1}^{c+1} \sum_{k=i+1}^{c+1} dev_{ik}^2 \quad (9)$$

$dev_{ik}^2$  is the deviation between the two centers  $c_i$  and  $c_k$ . Its value is determined by the exponent of the weight vectors centers  $\omega$ .  $\omega$  defines the fuzziness of the separation part:

$$dev_{ik}^2 = \mu_{ik}^{\omega} \|z_k - z_i\| \quad (10)$$

$[z_1, z_2, \dots, z_i, z_{i+1}, \dots, z_c]$  is the transposed matrix vector centers and the vector  $\bar{z}$ , that is the average of  $c$  centers.

$\mu_{ik}$  is the membership degree of  $z_k$  to the center  $z_i$ , its formula is:

$$\mu_{ik} = \left[ \sum_{l=i+1}^{c+1} \left( \frac{\|z_k - z_l\|}{\|z_k - z_i\|} \right)^{m-1} \right]^{-1}, \quad 1 \leq i \leq c+1, 1 \leq k \leq c+1, k \neq i \quad (11)$$

The index consists of a part of overall compactness and a fuzzy separation measure combining information on the data and the adhesion function. The overall compactness describes changes in class looking at the overall distribution of classes. The separation is based on the deviation between pairs of fuzzy centers. The performance of the index was examined by taking into account the two design parameters, namely the exponent of the fuzzy exponent  $m$  and the weight of the fuzzy separation  $\omega$ .

### 2.3 Partition coefficient and exponential separation

Proposed by Yang and Lung [10], this index detects a noise points in the dataset and eliminates a parasites. This index is type summation, while the remaining indices are of the type report.

This algorithm has a validity index for fuzzy clustering called Partition Coefficient And Exponential Separation (PCAES), It uses the factors of a normal class coefficient and exponential separation measure for each classification, and then combines these two factors to create the index PCAES. A consideration involving measures of compactness and separation for each classification provides various merits of validity of classes. Unlike other indices, the measure of validity given in PCAES can give another point of view in a noisy environment. For each class, we can measure the potential for it to be identified. Under this coefficient, a noisy point not offers enough opportunities to be an interesting class.

PCAES find an optimal evaluation of the number of classes, and provides more information about the data structure in a noise environment. PCAES formula is:

$$PCAES = \sum_{i=1}^c PCAES_i \quad (12)$$

$PCAES_i$  is the index of class  $i$ , its formula is:

$$PCAES_i = \sum_{j=1}^{n_i} \frac{\mu_{ij}^2}{n_i} \exp\left(-\frac{\min_k \|v_i - v_k\|^2}{\beta_T}\right) \quad (13)$$

with  $\sum_{j=1}^{n_i} \frac{\mu_{ij}^2}{n_i}$  is measurement of relative compactness of the class  $i$  compared to the most compact class that having value  $\mu_{ik}$ .

$\mu_{ik} = \min_{l \neq i} \left\{ \sum_{j=1}^n \mu_{ij}^2 \right\}$  is a compactness value of the most compact class,

$\exp\left(-\min\{\|a_i - a_k\|^2\} / \beta_T\right)$  is the separation measure of the class  $i$  with respect to  $\beta_T$ ,

$\beta_T = \frac{\sum_{i=1}^c \|v_i - \bar{v}\|^2}{c}$  is the total average measurement of the separation of  $c$  classes.

The compactness value belongs to the interval  $[0; 1]$ . The exponential separation function for class  $i$  measures the distance between class  $i$  and its closest neighbor class. This exponential measure is similar to the separation function defined by the XB index. Moreover, we consider the average measure of distance for all classes. Taking the exponential function to the separation measure in the interval  $[0; 1]$ .

With the compactness and separation for each class, the value PCAES<sub>i</sub> is calculated, which is the validity index of class i. The PCAES<sub>i</sub> could detect each class with two measures: a normal class coefficient and exponential separation. The great value of PCAES<sub>i</sub> means that class i is internally compact and separate from others (c - 1) groups. The small value of PCAES<sub>i</sub> indicates that the class i is not a identified cluster. The validity index of PCAES(c) is defined by summing all PCAES<sub>i</sub> to measure the compactness and separation of the data structure, as:

$$-c \leq PCAES(c) \leq c \quad (14)$$

The great value of PCAES(c) means that each of these classes is compact and separate from other classes. The small value means that some of these classes are not compact or separated from other groups. Moreover, the maximum PCAES(c), regarding c, could be used to detect the data structure with a compact class and well separated classes.

### 3. New Index

Fuzzy validity indices are more used than exact index, as their application in the exact domains is possible the same level as in the fuzzy domains. In fact, applied to FCM, the fuzzy indices imply the membership degrees of points to classes of the partition found, and considering 0 or 1. In contrast to other indices, the measure of validity proposed in PCAES can give a different view in a noisy environment. For each class, we can measure the potential for it to be identified. So using this index to determine the number of class, we developed a new version of a validity index named: Modified Partition Coefficient And Exponential Separation.

According to Xie and Beni [7], fuzzy compactness is given by the distance between the points  $x_j$  with membership degree  $\mu_{ij}$  and the center  $v_i$ . Also, according to Fukuyama-Sugeno [5] separation is of the form:

$$\sum_{i=1}^c \sum_{j=1}^n \mu_{ij}^2 d^2(v_i, v_j) \quad (15)$$

So the new formula is

$$MPCAES = Comp - Sep \quad (16)$$

The compactness value is :

$$Comp = \frac{\sum_{i=1}^c \sum_{j=1}^n \mu_{ij}^2 d^2(v_i, v_j)}{Comp_{min}} \quad (17)$$

The separation value is :

$$Sep = \exp \left( \frac{\frac{\sum_{i=1}^c \sum_{j=1}^n \mu_{ij}^2 \times \sum_{k=1}^c \sum_{l=1}^n \mu_{kl}^2 d^2(v_i, v_k)}{\sum_{j=1}^n \mu_{ij}^2 + \sum_{k=1}^c \sum_{l=1}^n \mu_{kl}^2}}{Sep_{max}} \right) \quad (18)$$

The most compact class is the class that has the minimum compactness, it formula is:

$$Comp_{min} = \min_{1 \leq i \leq c} \sum_{j=1}^n \mu_{ij}^2 d^2(v_i, v_j) \quad (19)$$

The maximum separation is the following:

$$Sep_{max} = \max_{1 \leq i \leq c} \frac{\sum_{j=1}^n \mu_{ij}^2 \times \sum_{k=1}^c \sum_{l=1}^n \mu_{kl}^2 d^2(v_i, v_k)}{\sum_{j=1}^n \mu_{ij}^2 + \sum_{k=1}^c \sum_{l=1}^n \mu_{kl}^2} \quad (20)$$

This fuzzy form will be used to apply in FCM algorithm by integrating the membership matrix U. We proceed testing MPCAES, and this by applying the FCM algorithm to the domain of data chosen for different values of c. In the experimental study of the next section, we provide an analysis of indices of validity resulting from the theoretical study of their implementation in the fuzzy and exact domains, and their application to FCM and HCM algorithms.

### 4. Some results

In this section, we evaluate the performance of studied validity indices, not only for the purpose of determining the optimal number of classes, but also to validate the structure of domains with different aspects: noise and overlapping.

**Example 1:** We will start with a two-dimensional synthetic base in Figure 2. The base is composed of three well-identified classes (optimal c = 3), it is used to verify the proper functioning of CVIs in a clean environment.

According to Figure 2, PCAES gave as optimal number of Class 5 and MPCAES gave 3 as the optimal number of class. So our index has determined the exact value of the number of classes.

**Example 2:** The experiments are applied to the indices already described in the theoretical study:

- PCAES: evaluates the noise aspect, and gets its optimum at the maximum value for different numbers of classes.
- SVI and XBI: promote measures of compactness and separation. These indices lead to optimal partitions to their minimum values.

We considered, to better test the noise aspect, the base shown in Figure 3, dataset on three well-identified classes without parasites points (a). We subsequently introduced at two levels noisy points: noise Level 1 (b) and noise level 2 (c).

To build quality indices studied in relation to the overlapping aspect, we took over the Figure 3 (a) not overlapped. We applied FCM (c = 2...12), and after the close of the three classes so that they reach different levels of overlap shown in Figure 4. We used the Dataset by touching a degree of overlap between classes combined



with noise points around these classes; it is shown in Figure 5.

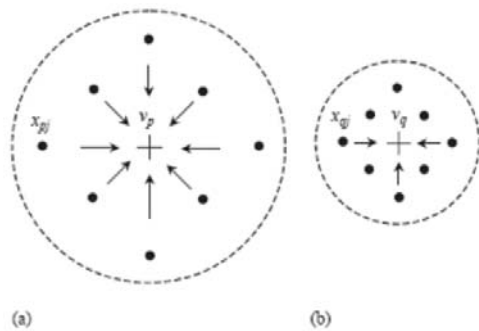


Fig. 1 Two fuzzy classes with similar values of compactness.

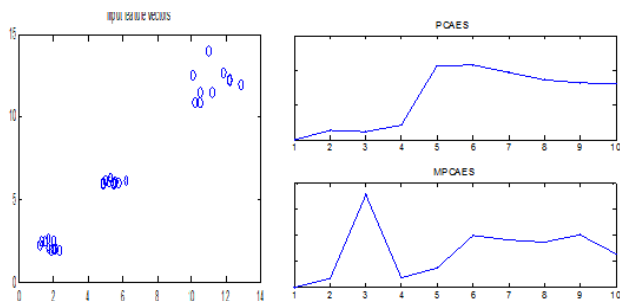


Fig. 2 Dataset two-dimensional synthetic with 30 points.

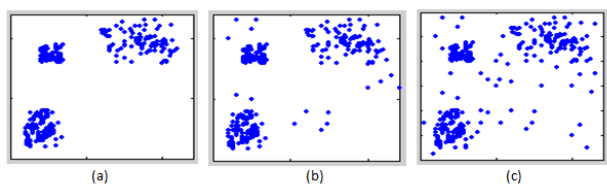


Fig. 3 Dataset base3 two-dimensional synthetic with 3classes: no-noisy base (a) noisy base level1 (b) noisy base level2 (c).

In the first level of noise introduced into the base3 in Figure 3.(b), the performance of PCAES and MPCAES is present to remove the noise points and consider only three optimal classes. SVI and XBI indices give 4 and 5 classes for optimal partitions, ignoring the noise present in the domain. In an interpretation of aspect of noise level 2 in Figure 3.(c), single index PCAES and MPCAES have kept the optimal partition composed of three classes: optimal  $c = 3$ , other indices have classified groups of parasitic points as valid classes (optimal  $c$  is 4 to 7). For data containing points of overlap level 1 and 2, shown in Figure 4., PCAES can provide a satisfactory result, but MPCAES gives better result, and this because of the

integration of degrees of membership in the compactness and separation. Other values indices found at take their minimum  $c = 2$ , which shows the limit in the existence of inter-class overlap in the data environment.

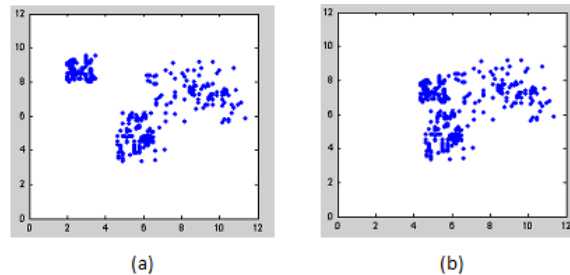


Fig. 4 Dataset base3, two-dimensional synthetic with 3 classes: overlapped base level 1 (a), overlapped base level 2 (b).

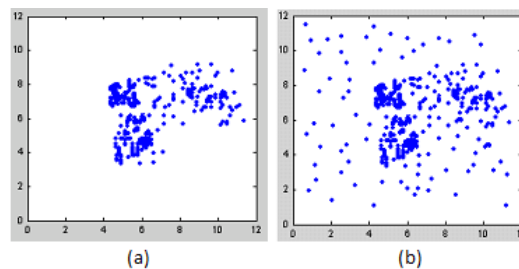


Fig. 5 Dataset base3, two-dimensional synthetic with 3 classes: overlapped base level 2(a), overlapped and noisy base level 2 (b).

We used the base3 by touching a degree of overlap between classes combined with parasites points around these classes; it is shown in Figure 5.6. A first observation on the behavior of indices resides in its failure to reach their optimum at  $c = 3$ .

We used the real base IRIS, also the MPCAES index opted for the best score by solving satisfactorily the problems of classification. The results of MPCAES index proposed in a number of domains have proven effective in noisy and overlap environments. However, all the indices presented in this study depend on the results of the FCM algorithm.

## 5. Conclusions

In this paper, we reviewed several validity indexes and then proposed a new validity index, called Modified Partition Coefficient And Exponential Separation, which is developed to obtain optimal partition. Moreover, we conducted extensive comparisons of the mentioned indices in conjunction with the FCM algorithm on a number of

widely used data sets. These results prove that our new index (MPCAES) provides the majority of cases the value of the desired classes.

Table 1: Number of classes found with the fuzzy indices

Base	SVI	XBI	PCAES	MPCAES
proper base3	3	3	3	3
noisy base3 level 1	4	4	3	3
noisy base3 level 2	5	4	3	3
Overlap base3 level 1	2	2	2	3
Overlap base3 level 2	2	2	2	3
Overlap and noisy base3 level 1	4	5	4	3
IRIS	3	2	3	3

## References

- [1] J. C. Bezdek, "Cluster Validity with fuzzy sets", J. Cybernetics, Vol.3,1974, pp. 58-73.
- [2] J. C. Bezdek and J. C. Dunn, "Optimal fuzzy partitions: A heuristic for estimating the parameters in a mixture of normal distributions", IEEE Transactions on Computers, Vol. 24, No. 8, 1975, pp. 835-838.
- [3] J. C. Bezdek and P.F. Castelaz, "Prototype Classification and Feature Selection With Fuzzy Sets", IEEE Transactions Syst Man Cybern, VOL. 2,1997, pp. 87-92.
- [4] R. Gunderson, "Applications of fuzzy ISODATA algorithms to startracker printing systems", Triannual world IFAC, 1978.
- [5] Y. Fukyama and T. Sugeno, "A new method of choosing the number of clusters for the fuzzy c-means method", in Proceedings of 5th Fuzzy System Symposium,1989, pp. 247-250.
- [6] I. Gath and A. B. Geva, "Unsupervised Optimal Fuzzy Clustering", IEEE Trans. Pattern Anal. Machine Intell., Vol. 7, 1989, pp. 773-781.
- [7] L. Xie and G. Beni, "A validity measure for fuzzy clustering, IEEE Trans PAMI, Vol. 13, No 8, 1991, pp. 841-847.
- [8] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 1, No 2, 1979, pp. 224-227.
- [9] G. E. Tsekouras and H. Sarimveis, "A new approach for measuring the validity of the fuzzy c-means algorithm", Advances in Engineering Software, Vol. 35, 2004, pp. 567-575.
- [10] K. Lung and M. S. Yang. "A cluster validity index for fuzzy clustering", Pattern Recognition Letters, Vol. 26, No 9, 2005, pp. 1275-1291.
- [11] D.W. Kim, K. H. Lee and D. Lee. "On cluster validity index for estimation of the optimal number of fuzzy clusters", Pattern Recognition, Vol.37 No. 10, 2004, pp. 2009-2025.

**Mohamed Fadhel SAAD** is currently a research staff member at Research Group on Intelligent Machines (REGIM). He is a Master Technologist to Higher Institute of Technological Studies of Gafsa. He received the Aggregation Specialty: Applied data processing to the management September 2000. In 2003 he started his Ph.D. studies at ENIS, University of Sfax, Tunisia. His main research interests are clustering, pattern recognition and fuzzy logics. He is a student member of the Institute of Electrical and Electronics Engineers (IEEE).

**Adel M. Alimi** received his Ph.D. degree in Electrical Engineering from Polytechnic school of Montréal, Canada, September 1995. He received this Habilitation To manage research (hdr) in Electric Engineering, option industrial computing, ENIS, University of Sfax, Tunisia. He is currently a Professor of industrial data processing at ENIS, University of Sfax, Tunisia since December 2006. His research interests include intelligent techniques, intelligent recognition of shapes and the manuscript, intelligent systems architecture and intelligent analysis of data. He has published prolifically in refereed journals, conferences, and workshops. He has served regularly in the organization committees and the program committees of many international conferences and workshops, and has also been a reviewer for the leading academic journals in his fields. He is a senior member of the Institute of Electrical and Electronics Engineers (IEEE), president of Tunisia Chapter of the IEEE Computer Society since 2010, advisor of the National Engineering School of Sfax Student Branch Chapter of the IEEE Robotics and Automation Society since 2010, advisor of the National Engineering School of Sfax Student Branch Chapter of the IEEE Computational Intelligence Society since 2010.

# An Accurate Arabic Root-Based Lemmatizer for Information Retrieval Purposes

Tarek El-Shishtawy<sup>1</sup> and Fatma El-Ghannam<sup>2</sup>

<sup>1</sup> Benha University  
Cairo, Faculty of Engineering (Shoubra) , Egypt

<sup>2</sup> Electronics Research Institute  
Cairo, Egypt

## Abstract

In spite of its robust syntax, semantic cohesion, and less ambiguity, lemma level analysis and generation does not yet focused in Arabic NLP literatures. In the current research, we propose the first non-statistical accurate Arabic lemmatizer algorithm that is suitable for information retrieval (IR) systems. The proposed lemmatizer makes use of different Arabic language knowledge resources to generate accurate lemma form and its relevant features that support IR purposes. As a POS tagger, the experimental results show that, the proposed algorithm achieves a maximum accuracy of 94.8%. For first seen documents, an accuracy of 89.15% is achieved, compared to 76.7% of up to date Stanford accurate Arabic model, for the same, dataset.

**Keywords:** Arabic NLP, Information Retrieval, Arabic Lemmatizer, POS tagger.

## 1. Introduction

In the field of NLP, lemmatization refers to the process of relating a given textual item to the actual lexical or grammatical morpheme [9]. Both of the word representation granularity level and its extracted morpho-syntactic features directly affect the performance of Information Retrieval (IR), Machine Translation, Summarization and keyphrase extraction systems. In semitic languages, such as Arabic, this is not an easy task due to their highly inflectional properties.

The granularity level is determined by clustering different words into groups, which shares the same root, stem or lemma. While many IR researchers consider the root level as the basic group others raise the importance of stem level for improving the effectiveness of IR systems. Lemma level analysis and generation does not get yet much focus in spite of its robust syntax, semantic cohesion, and less ambiguity

Lemma refers to the set of all word forms that have the same meaning, and hence capture semantic similarities between words. Recent researches in Arabic IR systems show the need of representing Arabic words at lemma

level for many applications, including statistical machine translation [10,13], keyphrase extraction [12], indexing and classification [17]. Lemmatization is relatively new topic for Arabic language processing, and hence only few researches focused directly on automatic lemma extraction from Arabic texts. Light stemmers and supervised learning approaches are the two main approaches for POS tagging and lemma extraction.

In spite of their limited accuracy, light stemmers and light lemmatizers introduce many useful techniques for disambiguating word category with minimum resources, which make them attractive to IR purposes. However, light stemmers fail in many cases to group related words [23], since there are no roots or stems to verify with. For example, it fails to conflate forms such as broken (irregular) plurals for nouns and adjectives with their singular forms, and past tense verbs do not get conflated with their present tense forms, because they retain some affixes and internal differences.

On the other hand, statistical supervised learning approaches present the best published accuracy as POS taggers. The cost of expanding language coverage is a major problem in supervised learning approaches. In closed learning methodologies, it is not an easy task to add new entries, since the whole model has to be retrained for these new entries. It is noted that [31], tagging accuracy for statistical approaches markedly decreases for previously unseen words. In our experimental results, the accuracy of -up to date- Stanford learning model for Arabic language was found to drop from its 96.86% best published result to only 76.7% for documents in education domain.

In this paper, we present a different approach that depends on Arabic language knowledge to disambiguate word category. We are motivated by the notion that the tagger performance can be improved by expanding the knowledge sources available to the tagger. In fact, Arabic language specialists are still able to disambiguate

unknown words (e.g., newly added Arabized words) by different language knowledge resources. In this paper, the proposed lemmatizer uses word patterns, roots, syntactic and morphological basic rules, to reduce Arabic words into their lemma canonical form. In the proposed approach, the extraction process is augmented with auxiliary dictionaries for words that are expected to fail in analysis with rules. The proposed approach extracts also the lexical category (POS) of the input word. To test the effectiveness of the proposed approach, we compared the results by up to date Stanford Arabic accurate tagger model. The results show that in all cases, better accuracy is obtained.

The rest of this paper is organized as follows: section (2) gives a brief description of related works in Arabic stemming techniques. The basic features of the algorithm are presented in section 3. The proposed algorithm has two main phases; POS tagging phase and lemma generation phase. The two phases are reviewed in sections 4 and 5. We present in section 4 the methodology and tools for building the proposed light lemmatizer. Section 6 gives performance analysis of the proposed algorithm. Finally, section 7 concluded the results.

## 2. Related works

Arabic is very rich in categorizing words, and hence, numerous stemming techniques have been developed for morphological analysis and POS tagging. Stemming and lemmatization shares a common purpose of reducing words to an acceptable abstract form, suitable for NLP applications. Previous works have presented important approaches that reflect different points of views to the problem. This includes debated word representation level, direct versus rule based lexicons, light stemmers versus accurate supervised learning approaches. The following subsections review these main trends and their relation to IR purposes.

### 2.1 Word representation level

There is no general agreement of the representation level of Arabic words in IR systems. Two levels have been debated; root level and stem level. Motivated by the power of Arabic roots, the first approach represents words in their root forms. Historically, a root was the entry to traditional Arabic lexicons. Almost, every word in Arabic is originated from a root. Many researchers [6,8], have emphasized that the use of Arabic roots as indexing terms substantially improves the information retrieval effectiveness over the use of stems.

However, several researchers [1,10] criticized this approach, and based their representation on stems. A stem is the least marked form of a word that is the uninflected word without suffixes, prefixes. Stem-based algorithms,

remove prefixes and suffixes from Arabic word until it matches one of the Arabic patterns, or generate such patterns from a root. Dichy and Fargaly [10], suggested stem level to be the basis of lexical entries of morpho-syntactic and semantic information. Also, Attia [5], concluded that using the stem as base form is far less complex in developing and maintaining, less ambiguous, and more suitable for syntactic parsers that aim at translation.

The main problem in selecting a root as a standard representation level in information retrieval systems is the over-semantic classification. Many words that do not have similar semantic interpretations are grouped into same root. For example, Arabic words ("مكتبة" library M123H), ("كاتب" writer 1A23) and ("كتاب" book 12A3) originate from the same root ("كتب" 123), while having different semantic senses. Thus, using the root-based algorithms in information retrieval may increase the word ambiguities.

On the other hand, stem level suffers from under-semantic classification. Stem pattern may exclude many similar words sharing the same semantic properties. For example, Arabic broken plurals have stem patterns which differs from their singular patterns. Stem-based representation, cannot detect the syntactic similarity between (طائر bird) and (طيور birds), since they have different stem patterns (1A23 and 12O3).

### 2.2 Direct and Rule-Based Lexicons

Direct access lexicons produce fewer errors, as they store a complete set of possible words, with their morpho-syntactic features. Early work on Arabic stemming used manually constructed dictionaries, till now this approach is still widely used. Al-Kharashi and Evens [2], worked with small text collections, for which they manually built dictionaries of roots and stems for each word to be analyzed. Buckwalter developed an Arabic Morphological Analyzer [7], as a set of lexicons of Arabic stems, prefixes, and suffixes, with truth tables indicating legal combinations. The main problem of direct access dictionary based approach is the dictionary size and the cost of maintenance.

In contrast, to direct accessed lexicons, other types of analyzers reduce the size of the corpus or dictionary by getting benefits from the derivational and inflectional productiveness of Arabic. Using morphological rules nouns and verb stems are derived from a few thousand roots by infixing. Such systems attempt to find the root, or any number of possible roots for each word. Xerox Arabic Morphological Analysis and Generation [6], is one of the superior rule based - root based systems. The system includes 4,930 roots and 400 patterns, effectively generating 90,000 stems.

Another superior open source root-based stemmer is the Khoja's stemmer [21]. A comparative study for three Arabic morphological analyzers and stemmers (Khoja stemmer, Buckwalter analyzer, and Tri-literal Root Extraction Algorithm), shows that Khoja stemmer achieves the highest accuracy [29].

The main criticisms to rule based approach are its coverage limitations, and abundance of many forms to reduce the ambiguity. For example, some of the most closely related forms such as singular and plural nouns are irregular, and are not related by simple affixing. For information retrieval, this abundance of forms means a greater likelihood of mismatch between the form of a word in a query and the forms found in documents relevant to the query. Another problem with rule based approach is that the generative power of rules makes it possible to produce forms that are ambiguous or unknown in the language.

### 2.3 Light Stemmer Approaches

Many statistical based approaches of IR systems rely on surface form of the word. This was the motivation for light stemming approaches. Light stemmers, are not an aggressive practice as the root-based algorithms. The aim of the light stemmers is not to produce the linguistic root of a given Arabic surface form, rather it is to remove the most frequent suffixes and prefixes [18]. Larkey's light stemmer [23], removes blindly the most frequent suffixes and prefixes from Arabic surface word given to produce the stem.

Hammouda et.al. [17], presented an approach to generate Arabic lemma for indexing purposes. The algorithm was based on removing suffixes, prefixes and all vowels from Arabic words without checking the validity of generated lemma. A similar heuristic approach was presented by Al-Shammari et.al. [3]. Their algorithm searches text words, hoping to find one of predetermined specific stop words, to differentiate between verbs and nouns and hence selects the appropriate stemming process. Both of the two algorithms can be categorized as light lemmatizers,

In spite of their limited accuracy, light stemmers and limmatizers introduce many useful techniques for disambiguating word category with minimum resources, which makes them attractive to IR purposes. However, light stemmers does not guarantee that the extracted lemma is a real word in Arabic language and suffers from limited extracted features. Moreover, light stemmers fail in many cases to group related words, since there are no roots or stems to verify with. For example broken plurals do not get conflated with their singular forms, and past tense verbs do not get conflated with their present tense forms, because they have internal differences in word structure.

### 2.4 Supervised Learning Approaches

Accurate analyzers cannot rely only on dictionary to disambiguate word categories. Many stems with different POS tags can be produced as possible stems of a given word. This problem exists in most languages; however it is serious in Arabic. For example, the word (عَلَّمَ teach) and (علم science) and (علم flag) are entries for same written Arabic word, as Arabic words are not normally written in their diacritic forms.

Many researchers deal with lemma extraction as a supervised learning problem. In general, these approaches give the best known accuracy as a POS tagger. The attempts started by Hajić [16], who built a training model to predict each word feature separately. Hajić used a direct dictionary as a source of possible morphological analyses (and hence tags) for an inflected word form. Statistical successful systems were presented by Ryan et.al. [28] for morphological disambiguation, Ibrahim et.al. [19], for mining parallel corpora to extract English-Arabic lemma pairs. The basic idea of these researches is to extract all possible analysis, and corresponding features, for each single word using Buckwalter Arabic Morphological Analyzer (BAMA). Features are then fed to a classifier, which is trained on data from Arabic Tree Bank (ATB), to disambiguate the word category. Correct features including lemma form are then extracted from a BAMA lexicon. The approach yields excellent results for previously seen words. Stanford Arabic POS tagger is another supervised system. It depends on different trained models for many languages including Arabic. The accurate model for Arabic, was trained on whole ATB for maximum entropy [30].

## 3. Features of the Proposed approach

Although there are many approaches for stemming Arabic, no single approach is considered as a standard IR-oriented stemming algorithm. In the current research we aim at building a lemmatizer with minimum sufficient resources for IR purposes. Lemmatizer transforms inflected word form to its dictionary lemma look-up form. For nouns and adjectives, lemma form is the singular indefinite (masculine if possible) form, and for verbs, it is the perfective third person masculine singular form.

The basic idea is to collect more information about the word to be stemmed and its context to generate more accurate word features including its POS tags. The system exploits Arabic language knowledge in terms of roots, patterns, affixes, and a set of morpho-syntactic rules to generate accurate lemma form and its relevant morpho-syntactic features that support information retrieval purposes. Morpho-syntactic features are required also, to capture the important semantic senses of the language.

Inflected languages such as Arabic, Finnish, Turkish, and Hungarian typically express meanings through morphological affixation. In highly Inflected languages plural, possessive relations, grammatical cases, and verb tenses, verb voice, and aspects which in English would be expressed with syntactic structures, are characteristically represented with morphological affixation [25]. Collected information about words also include word pattern. Arabic stem-patterns have interesting semantic features that give rise to senses of words. For example, syntactic patterns can recognize a given word as being the agent of an action, the instrument of that action, or the place at which the event occurs.

To implement our approach, the following features are considered:

- The proposed approach gets benefits of the power and generality of rule-based stemmers, and the accuracy of dictionary based approaches, which deals efficiently with cases of irregularity between singular and plural nouns, and proper nouns. Limited sized auxiliary dictionaries are used to enhance the performance of the lemmatizer.
- Relying only on lexicons may lead to much ambiguous word analysis possibilities. Therefore, Arabic context morpho-syntactic rules are used to expect the correct word category, and then verified. For example, the algorithm uses the word pattern and the category of its previous word for identifying the current word.
- In the current system, morphological and syntactic rules are adopted to reduce Arabic word into its abstract lemma form. Lemma is proved to be the smallest form that captures all semantic features of the word, and more suitable for IR systems.
- In spite of its importance in Arabic constructs, adjective is not classified as a POS main category of almost all Arabic POS taggers algorithms. Actually, traditional Arabic does not include adjective as one of its main parts-of-speech. An adjective in Arabic is actually a noun that happens to describe something. Many IR systems require aligning Arabic constructs with other languages constructs. Also, many IR systems extract candidate word category sequences that aid summarization and keyphrase extraction. In the proposed system, simple rules are used to re-categorize nouns as adjectives.
- The proposed system do not identify only word prefixes, suffixes and infixes, but finds out the corresponding morpho-syntactic attributes suitable for IR purposes at the lemma level.

The proposed algorithm has two main phases;

- POS tagging phase,
- Lemma generation phase.

#### 4. POS tagging phase

There are many morphological analyzers for Arabic; some of them are available as an open source for research and evaluation, while the rest are proprietary commercial applications. Instead of "reinventing the wheel", we started our analysis phase implementation with the open source Khoja stemmer [21]. To achieve the proposed lemmatizer features, many modifications both in data and basic algorithm flow were necessary to add Arabic knowledge.

Khoja morphological analyzer is a root based stemmer, which removes possible infixes of a word, finds corresponding matched pattern, and extracts the word root without POS tags. The list of roots consists of 3800 trilateral and 900 quad literal roots. Also, Khoja system recognizes a list of 168 Arabic stop words.

In our implementation of the analysis phase, the algorithm relies on using different knowledge resources of Arabic language: prefixes, suffixes, patterns, and rules. Limited size auxiliary dictionaries are used to augment morphological and syntactic rules in recognizing words, and resolving their ambiguity. The dictionaries include only words that are expected to fail in tagging by rules. In most cases, the ambiguity is due to the absence of the short vowels in the electronic Arabic documents, or non templatic word stems. The basic algorithm outline is shown in figure (1).

```
For each word (WO) Do
  Begin word_block
    Search a word in proper noun dictionary
    If exists POS = N, with features set, exit word_block.
    Check the existence in closed set word dictionary
    If found, POS = article with features set, exit word_block;..
    For each affix -longest first- Do
      Begin affix_block
        If affix cannot be removed from W then exit affix_block;
        Remove affix to extract the (W) form
        Check if (W) matches a pattern (P) with root R
        If (P) exists
          Begin POS_block
            Apply POS identification rules;
            If rules failed POS =N;
            Apply syntactic rules to detect Adjectives
          End POS_block;
        End affix_block;
      End loop;
    End word_block;
  End loop
```

Fig. 1 Outline of the Proposed Algorithm.

In the first stage, the algorithm starts the analysis by checking closed set Arabic words. The total list include 346 Arabic closed set words categorized into 16 groups (eg., prepositions, conjunctions, adverbs, numerals, etc.). Proper noun dictionary is also scanned at earlier stage of the analysis as shown in figure (1). The algorithm basic flow removes the longest suffix and the longest prefix in turn. After every elimination process, the algorithm checks a list of 61 patterns, if matches a pattern, the root is extracted and verified by checking the list of 3829 tri-roots. Up to this stage, the output is the suffix, prefix, word pattern, and root.

The purpose of the second stage is to tag POS of the word and to extract the corresponding features. The features for nouns and adjectives are definite case, count, and gender. Verbs extracted features are tense and voice. Finally, particles have 16 different subcategories. POS tagging and feature extraction are completed through many levels. The following subsections describe each level.

#### 4.1 Identifying Nouns and Verbs

In Arabic language, some verbs or inflected nouns can have the same orthographic form due to absence of vowels. However, the algorithm exploits many techniques to disambiguate Arabic lexical categories. Words are identified at different levels through our algorithm as follows:-

- a) The first level occurs after recognizing Arabic closed set words. The existence of closed words such as ( بعد - إلى - أمام - فوق - .... ) suggests that the next word is a noun. Also, The existence of closed words such as ( لن - كلما - لم - سوف - عندما ) suggests that the next word is a verb.
- b) The second level is the syntax rules. For example, one rule states that if the previous word is a verb, the current word cannot be also a verb, since Arabic language does not permit two successive verbs to exist
- c) Third level occurs during morphological stemming. In the proposed algorithm, affixes are categorized into three classes: affixes used by nouns only, affixes used by verbs only, and those that are used by either nouns or verbs. The existence of prefixes such as ( كال ، ك ، ب ، ل ) indicate that the word is a noun. Suffixes such as ( و ، - ني - نهم - وا ) indicate that the word is verb. Word tag of the first two classes is well defined, while a word that has prefix and/ or suffix of the third class is still ambiguous.
- d) The fourth level is the pattern-level, as illustrated in next subsection.
- e) The remaining words after verbs identification are considered nouns.

#### 4.2 Pattern level POS tagging

In our work, patterns play an important role in recognizing lexical word category. As shown in table (1). We classify Arabic patterns into three classes:

Table 1: Verb, Noun, and General patterns

Pattern Class	Pattern	Form	Examples
Verb Patterns	انفعال	en123	انتبه
	استفعال	est123	استقام
Noun Patterns	مفعول	m12o1	مكتوب
	فعال	12a3	كتاب
	افتعال	e1t2a3	اكتساب
General Patterns	فاعل	1a23	N شاعر - V ساعد
	تفاعل	t1a23	N تضايرب - V تظاهر
	فعل	123	N كتب - V كتب

- 1- Verb Patterns: which are used for verbs only.
- 2- Noun Patterns: which are used for nouns only.
- 3- General Patterns: which may be used for verbs or nouns according to different vocalization and not-written diacritics

If the word pattern belongs to first or second classes, it is a straight forward task to tag it as a noun or a verb respectively. For example, the word "ضوابط" corresponds to the pattern "فواعل", and hence the POS feature of the word is set to noun.

Words that have patterns belonging to third class cannot be tagged unless having a dictionary. In our implementation, instead of storing all Arabic verb forms (abstract and augmented), we store only most common verbs that belong to third class patterns. The current dictionary includes 943 common verbs, belonging to third class.

#### 4.2 Identifying Adjectives

Traditionally, Arabic does not include adjective as one of its main POS. An adjective in Arabic is actually a noun that happens to describe something. Adjectives take almost all the morphological forms of nouns. Adjectives, for example, can be definite, and are inflected for case, number and gender. In this stage of analysis noun words are checked to see if it is actually an adjective. The algorithm uses a shallow level sentence structure for this stage. In the proposed lemmatizer, the Noun category of a word is changed to an adjective if the following conditions are met

- 1- The current word does not contain any prefix.
- 2- Its previous word is also a noun (or adjective) and has the same count and gender.
- 3- Both of current and previous words are definite or both of them are indefinite.

In our implementation, the feature definite is different from [DET] used by Ryan et.al. [28], which checks only the existence of definite determinant. For example, Arabic nouns with attached possessive pronouns are definite in spite of non-existence of definite determinant.

## 5. Lemma Generation phase

The main contribution of this work is the development of lemma generator that extracts the lemma form of an Arabic word. On a word form conflation scale, lemma representation lies slightly above the (minimum) stem level, and below the (maximum) root level. The purpose of the second phase of the lemmatizer algorithm is to generate the abstract lemma form of a word. The next two subsections describe the procedures used to generate verb's and noun's lemmas.

### 5.1 Generating Verb's lemma

Verb lemma is the perfective, 3rd person, singular verb form. In most cases, lemma is the same as the root form of a verb. For example, the word (يكتب) has same root and lemma form (كتب). In other cases, removing prefixes and suffixes is not enough to generate the lemma form. In our implementation, rules are applied at the pattern level to deal with such cases. Table (2) shows examples of the analysis. The table shows that lemma form may be different from root form of a verb, and it may be required to remove or substitute prefix of the word stem.

Table 2: Examples of differences between pattern and lemma forms for verbs

Initial Word	After removing	Root	Pattern	Pattern form	Lemma form
تنازل	تنازل	نزل	تفاعل	t1a23	تنازل
يستخرجون	يستخرج	خرج	يستعمل	yst123	استخرج
نحتاجهم	نحتاج	حوج	نفتعل	n1t23	احتاج
تندرج	تندرج	درج	تنفعل	tn123	اندرج

### 5.1 Generating Noun's lemma

Lemma form of noun (or adjective) is the singular indefinite form. In Arabic, there are two types of noun and adjective plural forms: regular plurals, and broken (irregular) plurals. Regular plural can be masculine plural and feminine plural.

**Regular Plural:** The lemma form of the masculine plural is generated simply by removing prefixes "ون" or "ين" from a noun form. Lemma singular form of feminine plural nouns has two cases; feminine or masculine single form. Removing the suffix "ات" is enough for masculine single form case. Feminine singular form requires, adding "ة" to indicate its feminine nature. In our implementation, a dictionary is used to store feminine single noun forms.

The algorithm checks the noun word in the dictionary of the feminine words, if exists a character 'ة' is added.

Also, feminine have suffixes composed of 'ت' and attached pronoun (e.g., words (معالجتها، وظيفتك), requires suffix substitution when generating the lemma form. The rule is to replace the end character 'ت' with character 'ة', after removing the attached pronoun suffixes (lemma form will be (وظيفة and معالجة).

**Broken Plural nouns:** Another problem with nouns and adjectives lemma generation is the issue of broken plurals. The term was chosen to indicate that the base form of the nouns is broken either by removing one or more letters, adding one or more letters, changing vocalization or a combination of these. There are about 27 pattern forms for the broken plural [31], and in many cases, to generate the singular form, there are a lot of probabilities for the singular form pattern. For example if the broken plural pattern is (فعلاء) the singular form pattern may be one of the two patterns (فاعل، فعيل). Also, the broken plural words (كرماء، جهلاء) each of them has a pattern (فعلاء), and have different single forms (جاهل، كريم), which corresponds to the patterns (فاعل، فعيل) respectively, and there is no rule to determine exactly which of them is correct.

Therefore, it is very difficult to rely only on morphological rules without a dictionary to guess the lemma form of these broken plural ambiguous cases. In the proposed work, a dictionary is used to store only ambiguous cases, i.e., that have a lot of probabilities for the singular form.

## 6. Performance measurements

In order to evaluate the performance of the proposed system, two experiments were carried out on two different datasets. In the first experiment, a dataset is used to tune and evaluate the overall performance of the algorithm. In the second experiment, another dataset is used to compare the accuracy of the proposed approach and an up to date Stanford POS for unseen before documents. The following sub sections describe the details of each experiment

### 6.1 Experiment 1: Maximizing Performance

In the first experiment, a dataset was used to maximize the performance, and to measure the algorithm upper and lower accuracy bounds. The dataset contains 50 documents with 32,500 manually annotated words. The data are collected from different online resources and domains with focus on technical documents. It includes journal articles, technical reports, papers, and sections from text books. The first dataset is manually tagged to main POS (Nouns, Adjectives, Verbs, Articles, Proper Nouns, and Unknown). Unknown word category includes misspelling and non-Arabic words. Table (3), shows the



dataset POS tags as a percentage of the total words of the documents.

The purpose of the first experiment is to measure how much accuracy can be achieved with the proposed lemmatizer for document. Therefore, through the first experiment, some rules are rewritten, and words are added to maximize the performance. It is found that the proposed lemmatizer can achieve an accuracy of 94.8% of known documents. The efficiency is determined by the number of exact POS matches between manually tagged dataset words, and the proposed approach tagged words, divided by the total dataset words.

Table 3: POS tags for the first dataset

Word POS	% of total words
Nouns	50.0%
Verbs	8.4%
Adjectives	9.3%
Proper Nouns	3.7%
Closed System	25.9%
Unknown	2.7%

Through the experiment, different algorithm procedures are monitored to investigate their effect on the overall efficiency. The lower bound corresponds to running the algorithm with minimum resources. Through the experiment, it is found that using very simple procedures can lead to a minimum efficiency limit of 70% in case of ambiguous words. Storing only a look up dictionary of Arabic closed word set, simple definite prefixes, suffixes, and adjective rules, allow the algorithm to detect most of the nouns and adjectives successfully. The details of participation of each procedure are given below:

- 1) Looking up closed system Arabic words (particles), recognizes 25.9% of the total document words.
- 2) Using simple techniques for recognizing nouns leads to recognition of 75.4% of total nouns in the dataset, which corresponds to a 37.7% of the total document words. The details of these techniques are:
  - a. 28.2% of the document words are initially categorized as nouns by detecting the presence of definite prefixes "ال". Actually, this includes both definite nouns and definite adjectives:
    - i. Applying simple syntactic rules of adjective - only definite successive noun rule - allows tagging of 6.8% of total document words as adjectives. This corresponds to 73.1% of total adjectives in our dataset,
    - ii. The remaining 21.4% of definite document words are actual nouns.
  - b. 9.1% of the document words are succeeded to be categorized as nouns by detecting the presence of definite prefixes "لل".

- c. 2.5% of the document words are categorized as nouns by detecting the presence of prefixes "ك" and "ب".
- d. An additional 4.6% of the document words are categorized successfully as nouns by applying previous word category rules.

Table 4: Sample of results of the first experiment

Stanford Stemmer	Proposed Lemmatizer				English	Arabic word
	POS	R	P	L		
VBP	عدد	تفتعل	اعتمد	VV	It (female) depends	تعتمد
NOUN			معظم	particle	most	معظم
NN	بلد	فعلان	بلد	NNS	countries	بلدان
DTNN	علم	فاعل	عالم	DTNN	the world	العالم
DTNN			الآن	RB	now	الآن
IN			على	IN	on	على
NN	خدم	استفعال	استخدام	NN	use	استخدام
DTNN	نظم	افعل	نظام	DTNNS	the systems	الأنظمة
DTJJ	بني	مفعلة	مبنى	DTJJ	based	المبنية
IN			على	IN	on	على
DTNN	حاسب	فاعل	حاسب	DTNN	the computer	الحاسب
DTJJ			آلي	DTJJ	the automatic	الآلي
IN			في	IN	in	في
NN	نشأ	الفعال	انشاء	NN	building	إنشاء
NN	شغل	تفعل	تشغيل	NN +	and operating	وتشغيل
NN			صيانة	NN+	and maintenance	وصيانة
NN	شرع	مفاعيل	مشروع	NNS	projects	مشاريع
DTNN	بني	فعله	بنية	DTNN	the infra	البنية
DTJJ	سوس	الفعال	اساس	DTJJ	the basic	الاساسية
DTJJ	خوص	فعله	خاصة	DTJJ	the dedicated	الخاصة
NN			بها	particle	for it	بها
IN			في	IN	in	في
NN	خلف	مفتعل	مختلف	NN	different	مختلف
DTNNS	قطع	فعال	قطاع	DTNNS	the sectors	القطاعات
NN	مثل		مثل	NN	like	مثل
NNS	قطع	فعال	قطاع	JJ	sectors	قطاعات
DTNN	صنع	فعالة	صناعة	DTNN	the industry	الصناعة
NN	زرع	فعالة	زراعة	DTNN +	and the agriculture	والزراعة
NN			صحة	DTNN +	and the health	والصحة
NNP	علم	تفعل	تعليم	DTNN +	and the education	والتعليم
NNP	تجر	فعالة	تجارة	DFNN +	and the commerce	والتجارة
NNP	بنك	فعل	بنك	DTNNS +	and the banks	والبنوك
NNS	خدم		خدمة	DTNNS +	and the services	والخدمات
JJ .			غير	particle	And others	وغيرها
PUNC					PUNC	.

Hence, in this minimum resource experiment, approximately 70% of total Arabic document words can be tagged without use of any roots or patterns as resources. This is the minimum boundary of the algorithm, which is equivalent to light stemmers operation. However, this

minimum accuracy level is not accepted by many IR systems, since until now, no verbs are detected.

### 6.1 Experiment 2: Comparing Results

In the second experiment, we compare our root-based Arabic Lemmatizer output to Stanford POS Arabic tagger [30], with trained model (Arabic-accurate model dated 2011-09-14). The Stanford model was trained on the entire ATB p1-3. A second dataset is used through this experiment, which is a non-technical Arabic document collected for building Contemporary Arabic Corpus [4], The document contains 4020 Arabic words in education domain.

```
كانKAN/النظامDTNN/التعليميDTJJ/فيIN/العراقDTNPP/منIN/أكثرNN  
النظمDTNNS/تقدّمًاparticle/فيIN/العالمDTNN/العربيDTJJ/قبلADV  
عامNN/1990/بيدNN/أنparticle/هذاDEMO/النظامDTNN/تدهورVV  
تدهورًاVV/كبيرًاNN/نتيجةNN/الحروبDTNNS/التيCONJ/تورطVV  
فيهاparticle/النظامDTNN/السابقDTJJ/وماNN/أعقبهاJJ/منIN/فرضNN  
عقوباتNNS/دوليةJJ/علىIN/البلادDTNNS/مماparticle/أدخلهاNN/فيIN  
دائرةNN/الإهمالDTNN/والانعزالDTNN+/أورثNN/مشكلاتNNS/ضخمةJJ  
مماparticle/زالتVV/البلادDTNNS/تعانيهاNN/فيIN/الوقتDTNN  
الحاليparticle.وقدحرفتفاقتVV/الأوضاعDTNNS/نتيجةNN/أعمالNNS  
التدميرDTNN/والتهب+DTNN/والتعطيل+DTNN/لمؤسسات+NNS  
الدولةDTNN/والتيCONJ/وقعتVV/منذIN/شهرADV/مارسNN/2003مNUM  
فيIN/أعقابNNS/سقوطJJ/العاصمةDTNN/بغدادNNP+/وانهيار+NN  
النظامDTNN/السياسيDTJJ/ودخول+NN/القواتDTNNS/الأمريكيةunknow  
والبريطانيةunknow/للبلاد+DTNNS/ويأملVV/المجتمعDTNN/الدوليDTJJ  
بعدADV/أنparticle/يستتبunknow/الأمرDTNN/فيIN/العراقDTNPP  
ويؤولVV/الحكمDTNN/ومقاييد+NN/السلطةDTNN/لحكومة+NN/وطنيةJJ  
عراقية+DTNNS/منتخبةNN/أنparticle/يتحركVV/العراقDTNPP/بسرعةNN  
+ لإعادة+NN/بناءNN/النظامDTNN/التعليميDTJJ/وتأهيله+NN/وتجديده+NN
```

Fig. 2 Output of proposed lemmatizer, 15 mistakes were detected out of 110 words (non-technical document). Mistakes are shown bolded.

```
كان VBD/النظامDTNN/التعليميDTJJ/فيIN/العراقDTNPP/منIN/أكثرNN  
النظمDTNNS/تقدّمًاparticle/فيIN/العالمDTNN/العربيDTJJ/قبلADV  
عامNN/1990/بيدNN/أنparticle/هذاDEMO/النظامDTNN/تدهورVV  
تدهورًاVV/كبيرًاNN/نتيجةNN/الحروبDTNNS/التيCONJ/تورطVV  
فيهاparticle/النظامDTNN/السابقDTJJ/وماNN/أعقبهاJJ/منIN/فرضNN  
عقوباتNNS/دوليةJJ/علىIN/البلادDTNNS/مماparticle/أدخلهاNN/فيIN  
دائرةNN/الإهمالDTNN/والانعزالDTNN+/أورثNN/مشكلاتNNS/ضخمةJJ  
مماparticle/زالتVV/البلادDTNNS/تعانيهاNN/فيIN/الوقتDTNN  
الحاليparticle.وقدحرفتفاقتVV/الأوضاعDTNNS/نتيجةNN/أعمالNNS  
التدميرDTNN/والتهب+DTNN/والتعطيل+DTNN/لمؤسسات+NNS  
الدولةDTNN/،VBD/والتيCONJ/وقعتVV/منذIN/شهرADV/مارسNN  
2003مNUM/فيIN/أعقابNNS/سقوطNN/العاصمةDTNN/بغدادNNP+/وانهيار+NN  
النظامDTNN/السياسيDTJJ/ودخول+NN/القواتDTNNS/الأمريكيةDTJJ  
والبريطانيةNN/للبلادPUNC./ويأملNN/المجتمعDTNN/الدوليDTJJ  
بعدADV/أنNN/يستتبVBP/الأمرDTNN/فيIN/العراقDTNPP/ويؤول+NN  
الحكمDTNN/ومقاييد+NN/السلطةDTNN/لحكومة+NN/وطنيةJJ/منتخبةJJ  
أنVBD/يتحركVBP/العراقDTNPP/بسرعة+NN/لإعادة+NN/بناءNN  
النظامDTNN/التعليميDTJJ/وتأهيله+NNP./PUNC.وتجديده+NNP
```

Fig. 3 Output of Stanford Arabic stemmer, 28 mistakes were detected out of 110 words (non-technical document). Mistakes are shown bolded.

In our experiment on the unseen before dataset, the average accuracy of the proposed algorithm is 89.15% as a POS tagger, while it is 76.7% for Stanford Arabic accurate

model. Figure (2), and figure (3), show the output of both algorithms for the same sample text. In our algorithm, most of the errors encountered result from undiscovered verbs, proper nouns, and confusing nouns with adjectives, Stanford Arabic stemmers errors are mainly due to improper detection of Arabic broken plurals, and limited coverage of basic Arabic words.

In an error analysis of Stanford POS tagger [26], it is noted that 4.5% of errors are due to unknown word, where the tagger has to rely only on context features, and contexts are often ambiguous. Stanford tagger for Arabic complains real problems with categorizing broken plural nouns, definite nouns detection, and noun-verb confusion.

## 7. Conclusions

In this research we have presented an accurate algorithm for extracting lemma form of Arabic words and their morpho-syntactic features. The presented algorithm proves that accurate results for POS tagging, can be achieved when using inherent features and rules of Semitic languages like Arabic. It is shown that ambiguity can be resolved using metadata about patterns, roots, and infixes' indications. Analysis is aided with auxiliary dictionaries and syntax rules to produce a lemmatizer which outperforms existing up to date Arabic learning algorithms. .

## References

- [1] Al-Jlayl, M., Frieder O., 2002 . "On Arabic Search: Improving the Retrieval Effectiveness via Light Stemming Approach". In Proceedings of the 11th ACM International Conference on Information and Knowledge Management, Illinois Institute of Technology, New York: ACM Press, pp. 340-347.
- [2] Al-Kharashi L., and M.Evens "Comparing words, stems, and roots as index terms", in an Arabic information retrieval system. JASIS, 45 (8), 1994, pp. 548-560.
- [3] Al-Shammari E., and J. Lin, "A Novel Arabic Lemmatization Algorithm in Proceedings of the second workshop on Analytics for noisy unstructured text data, ACM, 2008.
- [4] Al-Sulaiti L., and E.Atwell, "Extending the Corpus of Contemporary Arabic". in Proceedings of Corpus Linguistics conference, University of Birmingham, UK, 2005.
- [5] Attia M., "An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modeling Finite State Networks", in The Challenge of Arabic for NLP/MT Conference. The British Computer Society, 2006.
- [6] Beesley, K. "Arabic finite-state morphological analysis and generation" ,In COLING-96: Proceedings of the 16th international, 1996. pp. 89-94.
- [7] Buckwalter, T. Qamus: Arabic lexicography. <http://members.aol.com/ArabicLexicons/>

- [8] Darwish, K., Doermann, D., Jones, R., Oard, D., and Rautiainen, M., 2001. "TREC-10 experiments at Maryland: CLIR and video" In TREC 2001. Gaithersburg: NIST.
- [9] Dicky J. "On lemmatization in Arabic. A formal definition of the Arabic entries of multilingual lexical databases". ACL 39th Annual Meeting, Workshop on Arabic Language Processing; Status and Prospect, Toulouse, 2001, 23-30.
- [10] Dicky J, and A.Fargaly, "Roots & Patterns vs. Stems plus Grammar-Lexis Specifications: on what basis should a multilingual lexical database centred on Arabic be built?", Proceedings of the MTSummit, New-Orleans,2003.
- [11] El Isbihani A., K.Shahram, O. Hermann, " Morpho-syntactic Arabic Preprocessing for Arabic-to-English Statistical Machine Translation" in Proceedings of the Workshop on Statistical Machine Translation, pages 15–2 New York City, June 2006.
- [12] El-Shishtawy T. and. Al-Sammak A. "Arabic Keyphrase Extraction using Linguistic knowledge and Machine Learning Techniques", Proceedings of the Second International Conference on Arabic Language Resources and Tools, The MEDAR Consortium, Cairo, Egypt. 2009.
- [13] Fedaghi A, Al-Anzi, F. S.,. "A new algorithm to generate Arabic root-pattern forms". In Proceedings of the 11th national computer conference. King Fahd University of Petroleum & Minerals, Dhahran, Saudi Arabia, 1989,pp. 391-400.
- [14] Habash N, Arabic Morphological Representations for Machine Translation in Representations of Arabic Morphology, Soudi A., Van Bosch A. (Ed), Neumann G. (Ed) (2007).
- [15] Habash, N, A.Soudi, and T.Buckwalter, On Arabic Transliteration.: Arabic Computational Morphology-Knowledge-based and Empirical Methods. Van Bosch and A. Soudi. 2007.
- [16] Hajic J., "Morphological tagging: Data vs. dictionaries". in 1st Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL'00), Seattle, WA, 2000.
- [17] Hammouda F., and A.Almarimi, "Heuristic Lemmatization for Arabic Texts Indexation and Classification" in Journal of Computer Science 6 (6): 660-665, 2010.
- [18] Hayder K., K.Al Ameer, O.Shaikha Al Ketbi, A.Amna A. Al Kaabi, Khadija S. Al Shebli, Naila F. Al Shamsi, Noura H. Al Nuaimi, Shaikha S. Al Muhairi "Arabic Light Stemmer: Anew Enhanced Approach" The Second International Conference on Innovations in Information Technology (IIT'05)
- [19] Ibrahim S. and N. Habash, "Automatic Extraction of Lemma-based Bilingual Dictionaries for Morphologically Rich Languages". in The Third Workshop on Computational Approaches to Arabic Script-based Languages at the Machine Translation Summit XII Ottawa, Ontario, Canada, 2009.
- [20] Khemakhem I, S.Jamoussi, A.Hamadou "Arabic morpho-syntactic feature disambiguation in a translation context" Proceedings of SSST-4, Fourth Workshop on Syntax and Structure in Statistical Translation, pages 61–65, Beijing, 2010.
- [21] Khoja, S., 1999. "Stemming Arabic Text". Lancaster, U.K., Computing Department, Lancaster University. www.comp.lancs.uk/computing/users/khoja/stemmer.ps.
- [22] Kibort A., "What are morphosyntactic features." <http://www.surrey.ac.uk/lis/smg/morphosyntacticfeatures.htm> 2007
- [23] Larkey L., and I.Ballesteros, "Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis", Proceedings of the 25th annual international ACM SIGIR conference, 2002.
- [24] Leah S. L.Larkey, and M. Connell, "Light Stemming for Arabic Information Retrieval"
- [25] Löfberg L., L.Archer, D.Piao, S.Rayson, P.Mcenery, T.Varantola, K.Juntunen "Porting an English semantic tagger to the Finnish language", In: Proceedings of the Corpus Linguistics 2003.
- [26] Manning C.. "Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics? In Alexander Gelbukh (ed.), Computational Linguistics and Intelligent Text Processing, 12th International Conference, Proceedings, Part I. Lecture Notes in Computer Science 6608, Springer, 2011
- [27] Pakray P., A. Gelbukh and S.Bandyopadhyay, "A Hybrid Textual Entailment System using Lexical and Syntactic Features", In The 9th IEEE International Conference on Cognitive Informatics, (ICCI 2010), pp. 291-295. 2010.
- [28] Ryan R, O.Rambow, N.Habash, M.Diab, and C.Rud , "Morphological Tagging, Diacritization, and Lemmatization Using Lexeme Models and Feature Ranking" in Proceedings of ACL-08: HLT, pages 117–120, Columbus, Ohio, USA, June 2008.
- [29] Sawalha M., and E.Atwell, "Comparative Evaluation of Arabic Language Morphological Analysers and Stemmers", Coling 2008: Companion volume – Posters and Demonstrations, Manchester, 2008, pages 107–110.
- [30] Stanford Natural Language Processing Group, Arabic-accurate.tagger, <http://nlp.stanford.edu/software/tagger.shtml>, 2011.
- [31] Toutanova K., and C.Manning, "Enriching the Knowledge Sources used in a Maximum Entropy Part-of-Speech tagger". In: Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora (EMNLP/VLC-2000) , Hong Kong 2000.
- [31] فؤاد نعمة المكتب العلمي للتأليف والترجمة، " ملخص قواعد اللغة العربية 18، 1973. الطبعة
- [32] مؤسسة الكتب الثقافية، " شذا العرف فى فن الصرف".1894، الشيخ احمد الحملاوي.
- Tarek El-Shishtawy** is a Professor assistant at Faculty of Engineering, Benha University, Egypt. He participated in many Arabic computational Linguistic projects. Large Scale Arabic annotated Corpus, 1995, was one of important projects for Egyptian Computer Society, and Academy of Scientific Research and Technology, He has many publications in Arabic Corpus, machine translation, Text, and data Mining.
- Fatma El-Ghannam** is a researcher assistance at Electronics Research Institute – Cairo, Egypt. She has great research interests in Arabic language generation and analysis. Currently, she's preparing for a Ph.D. degree in NLP.

# Synthesis of Quantum Multiplexer Circuits

Arijit Roy, Dibyendu Chatterjee and Subhasis Pal

Department of Electronics, West Bengal State University  
 Barasat, Kolkata, Pin 700 126, India

## Abstract

Combinational quantum circuits are essential for quantum computation; and quantum multiplexer circuit is one of the important combinational circuits. In this paper, we have presented the synthesis of quantum multiplexer circuit in detail. Instead of using functional blocks, we have used physically realizable quantum logic gates for synthesis of quantum multiplexers. In addition to this, our synthesis procedure shows that it is possible to construct quantum multiplexer circuit that can operate in the minimum dimension of the vector space and scalable according to linear nearest neighbor architecture. The detail functionality of the circuits along with the matrix formulations is presented.

**Keywords:** *Quantum Synthesis, Quantum Circuits, Quantum Multiplexer.*

## Nomenclature

Qbit	Quantum bit
Qgate	Quantum gate
QNOT	Quantum NOT gate
LNN	Linear nearest neighbor
MUX	Multiplexer
CMUX	Classical digital multiplexer
QMUX	Quantum multiplexer
$\mathcal{H}_n$	n-dimensional Hilbert space, $2^n$ -dimensional vector space (n=1, 2, 3, ... etc.)
CNOT	Quantum controlled-NOT gate with one control Qbit
$C^n$ NOT	Quantum controlled-NOT gate with n number of control Qbits (n>1)
SWAP	Quantum swap gate
CSWAP	Quantum swap gate with one control Qbit
$C^n$ SWAP	Quantum swap gate with n number of control Qbit (n>1)

## 1. Introduction

The pressure of fundamental limits on classical computation and the promise of exponential speedups from quantum mechanical effects are recently brought quantum circuits to a new dimensional attention of electronics community. As a result of which the quantum computation and information remain an attractive area of research in the last couple of decades. It is noticed that though wealth of knowledge in quantum mechanics is acquired; today development of quantum computer suffers

from many aspects. Synthesis of quantum circuits is one of the major challenges in the quantum information processing and in the development of the architecture of quantum computer. Though some basic quantum logic circuits and gates are demonstrated, the efficient functional blocks such as quantum flip-flop, register, multiplexer, demultiplexer, counter etc. have not demonstrated and investigated rigorously to produce efficient quantum circuit which can be constructed by physically realizable Qgate. This paper deals with one of the combinational functional blocks such as QMUX which can be constructed by physically realizable gate such as CNOT gate.

In quantum computation, the Qbits are counterpart of the classical bits. Unlike bits which are described by two constants (0 and 1) and manipulated using Boolean algebra, Qbits are described in terms of vectors, matrices and manipulated using other linear algebra. The Qbits are realized in Hilbert space ( $\mathcal{H}_1$ ) spanned by the orthogonal basis states  $|0\rangle$  and  $|1\rangle$ , i.e.

$$\mathcal{H}_1 = \text{span}_c \{|0\rangle, |1\rangle\} \quad (1)$$

A Qbit can be in a superposition state that combines  $|0\rangle$  and  $|1\rangle$ . The states,  $|0\rangle$  and  $|1\rangle$  are the vectors of the computational basis and the value of a Qbit can be any unit vector in the space they span (i.e. in  $\mathcal{H}_1$ ).

In addition to this, unlike the classical logic gate operation, the operation on Qbits must be reversible. The reversibility requirement of the operation on Qbits poses another challenge in the circuit synthesis. Both the logical and physical reversibility are the concern of any quantum circuits. If a circuit is logically reversible, then inputs can be constructed from the outputs of the circuit. For example, among the classical logic gates, NOT gate is the only reversible gate, but it is not a universal gate. While in quantum circuits, Fredkin gate, Toffoli gate (both having 3 inputs and 3 outputs) are the popular universal as well as reversible quantum gates. So in the cases, where the operation of a quantum circuit consists of many quantum operations, it is extremely important to check the reversibility of all the operations involved in that quantum circuit. Apart from these issues, Qbits cannot be copied using quantum wire in a similar way that we normally do

for classical circuits. Additionally, the number of inputs and outputs in any quantum circuit must be same.

The function of a MUX is to select one input among a group of inputs and pass the selected input to output of the circuit. Basically it consists of two types of inputs: one group is the data input and the other group is the select input and these select inputs decide which data input is to be selected to pass to the output. A classical “d:1 MUX” implies a MUX circuit with d number of data input and one output. A MUX circuit has numerous applications in information processing and communication.

Developing electronic functional block using another functional block is very common in electronics. For example, classical registers which are commonly composed of flip-flops. In such development, cost (of fabrication) and time taken for operation are mainly considered as efficiency of circuit. Recently, a few QMUX circuits are synthesized and presented [1,2].

The synthesis of QMUX using ternary quantum gates is also presented [2]. Since, the ternary quantum state is difficult to achieve and quite immature as a quantum effect, we have considered the most commonly used binary quantum state in the circuit synthesis procedure. Vivek et al. [3,4] and K. N. Patel et al. [5] presented many elements of the theory of quantum circuit to construct combinational circuits and we have extensively used their work in the synthesis of optimal QMUX.

In this work, we have shown that it is possible to develop QMUX circuit using physically realizable quantum logic gates. Open source software package ‘Octave’ is used as programming tool for this work. The operations involved in the proposed circuit are very basic in nature. We found that the number of operations and the cost of the proposed quantum circuit are optimum. The functionality of the circuit along with the reversibility requirement and matrix formulation are provided. The generalization of higher order QMUX synthesis is also presented.

## 2. Background

A combinational quantum logic circuit consists of quantum gates, interconnected by quantum wire carrying Qbits without fanout or feedback. Since, each quantum gate has the same number of inputs and outputs; any cut though the circuit crosses the same number of wires [3]. Quantum circuit operation is sequence of some quantum logic operations by some Qbits. A quantum wire is realized by a Qbit and corresponding matrix is a  $2 \times 2$  identity matrix.

On the other hand, a quantum logic gate is a closed-system evolution (or transformation) of the n Qbit state space  $\mathcal{H}_n$ , i.e.

$$\mathcal{H}_n = \text{span}_{\mathbb{C}}\{|q\rangle; q \text{ a bitstring of length } n\} = \text{span}_{\mathbb{C}}\{|q_1\rangle, |q_2\rangle, |q_3\rangle, |q_4\rangle, \dots, |q_{2^n}\rangle\} \quad (2a)$$

Where  $|q_i\rangle = |b_0 b_1 b_2 b_3 b_4 b_5 \dots b_{n-1}\rangle = |b_0\rangle|b_1\rangle|b_2\rangle|b_3\rangle \dots |b_n\rangle$  for each  $b_i \in \{0,1\}$ ;  $|b_0 b_1 b_2 b_3 b_4 b_5 \dots b_{n-1}\rangle$  is abbreviated as *bitstring state* and  $|q\rangle = |\text{no of bits}\rangle$ .

Here the arbitrary vector  $|\psi\rangle$  ( $|\psi\rangle \in \mathcal{H}_n$ ) can be written as  $|\psi\rangle = \alpha_1|000\dots00\dots000\rangle + \alpha_2|000\dots00\dots001\rangle + \alpha_3|000\dots00\dots010\rangle + \dots + \alpha_i|000\dots01\dots101\rangle + \dots + \alpha_n|111\dots11\dots111\rangle$  (2b)

$$|\psi\rangle = \sum_{q \in \mathbb{Q}^n} \alpha_i |q\rangle = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \vdots \\ \alpha_i \\ \vdots \\ \alpha_{2^n} \end{pmatrix} \quad (2c)$$

Where,  $\mathbb{Q}^n$  is the space of bitstring of length of n and  $|\alpha_i|^2$  is the probability of the state of  $i^{\text{th}}$  element. So a n Qbit space is  $2^n$  dimensional vector space and this can also be utilize as n Qbit register and a n Qbit circuit. Or in other words, no information is gained or lost during the transformation. Thus if  $|q\rangle$  is a state vector in  $\mathcal{H}_n$ , the operation of n Qbit quantum logic gate can be represented by

$$|q\rangle \rightarrow U|q\rangle \quad (2d)$$

Where, U is the  $2^n \times 2^n$  unitary matrix representing the gate operation.

Before we proceed to the synthesis of QMUX circuit, it is important to understand the effect of parallel and/or series combinations in quantum circuits and circuit elements used to construct the QMUX. We have used  $2 \times 2$  SWAP gate and multiple controlled SWAP gate in the synthesis of QMUX. For the purpose of the quantum cost calculation, the multiple controlled SWAP gate is decomposed in terms of CNOT gate, controlled V gate and controlled  $V^\dagger$ . In the following paragraphs, we have presented these aspects of quantum circuits.

### 2.1 Combination of Quantum Circuit Elements

Combinational circuits are important to build a functional block. In order to demonstrate the effect of various combinations (series and parallel) of quantum gates and quantum wire in quantum circuits, a circuit shown in Fig. 1 is considered. The equivalent circuit is also shown in this figure.

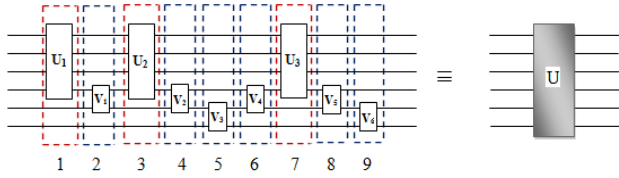


Fig. 1 A typical quantum logic circuits and its equivalence.

The circuit shown in Fig. 1 is a six Qbit quantum circuit and the circuit is composed of four Qbit and two Qbit quantum gates. The number of four and two Qbit Qgate in the circuit is 3 and 6 respectively. Note that the state of six Qbit circuit can be expressed by a vector in  $\mathcal{H}_{6(n=6)}$  (contains  $2^6$  column element), whereas the four Qbit and two Qbit gates are expressed by unitary operations on  $\mathcal{H}_{4(n=4)}$  and  $\mathcal{H}_{2(n=2)}$ . If U be the resultant unitary matrix representing the six Qbit circuit operation (the dimension of the U will be  $2^6 \times 2^6$ ), then

$$U = (I^{\otimes 4} \otimes V_6)(I^{\otimes 3} \otimes V_5 \otimes I)(U_3 \otimes I^{\otimes 2})(I^{\otimes 3} \otimes V_4 \otimes I)(I^{\otimes 4} \otimes V_3) \\ (I^{\otimes 3} \otimes V_2 \otimes I)(U_2 \otimes I^{\otimes 2})(I^{\otimes 3} \otimes V_1 \otimes I)(U_1 \otimes I^{\otimes 2}) \quad (3)$$

The number 1 to 9 at the top/bottom of the Equation (3) represents the individual operation of each block of circuit labeled by 1 to 9 in the circuit (see Fig. 1).

From the Equation (3) one can see that an individual operation can be represented by the tensor product (parallel combination of quantum wire and Qgate are represented by tensor product of corresponding unitary matrices of quantum circuit elements) of the corresponding space matrices while the linear combination of two individual operations is represented by the ordinary product of the individual space matrices. For example, consider the individual operation 1, which can be represented by the tensor product of  $U_1$  and  $I^{\otimes 2}$  i.e.  $(U_1 \otimes I^{\otimes 2})$ . Similarly, the operation 2 can be represented by  $(I^{\otimes 3} \otimes V_1 \otimes I)$ , while series combination of the operations 1 and 2 is the multiplication of their space matrices i.e.  $(I^{\otimes 3} \otimes V_1 \otimes I)(U_1 \otimes I^{\otimes 2})$ .

Actually six Qbits go through the operation according to the quantum circuit to produce some output state. Ordinary products of nine consecutive operations are performed by the circuit to produce some output states from the input states.

## 2.2 CNOT Gate

CNOT gate is one of the fundamental logic gates in quantum circuits and it operates in four ( $2^2$ ) dimensional space. This gate consists of two inputs: one is the control input (a Qbit) and the other one is the target Qbit. The

circuit and operational matrix  $U_{\text{CNOT}} (2^2 \times 2^2)$  of the CNOT gate is shown in Fig. 2. The operation of this gate can be written as  $|\psi\rangle \rightarrow U_{\text{CNOT}}|\psi\rangle \rightarrow |a, b \oplus a\rangle$ , where  $|\psi\rangle = |a\rangle|b\rangle = |a\rangle \otimes |b\rangle$ .

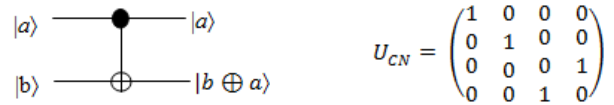


Fig. 2 Circuit of CNOT gate and its matrix. Here  $|a\rangle$  and  $|b\rangle$  are the control and target Qbit respectively.

A three-input controlled-controlled-NOT gate in which two inputs act as control Qbit and the rest one acts as target Qbit is known as Toffoli gate (or  $C^2$ NOT gate) and is a universal quantum gate. This gate operates in 8 ( $=2^3$ ) dimensional space. The circuit of the  $C^2$ NOT gate and its matrix  $U_T (2^3 \times 2^3)$  is shown in Fig. 3. The operation of this gate can be written as  $|\psi\rangle \rightarrow U_T|\psi\rangle = U_T|a, b, c\rangle = |a, b, (c \oplus a.b)\rangle$ .

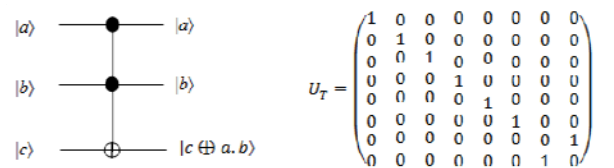


Fig. 3 Circuit of  $C^2$ NOT gate (or Toffoli gate) along with its matrix. In this circuit  $|a\rangle, |b\rangle$  are the two control Qbits and  $|c\rangle$  is the target Qbit.

Though the CNOT gate is not directly involved in the synthesis of QMUX circuits, we have presented the above emphasis on CNOT gate since, we have used SWAP gate to construct QMUX circuits and SWAP gate consists of CNOT gate. In addition to this, CNOT gate is also involved in quantum cost calculation.

## 2.3 SWAP Gate

It swaps the states of two Qbits. The operation of this gate can be decomposed into three CNOT operations. The circuit for SWAP gate and its matrix  $U_S (2^2 \times 2^2)$  is shown in Fig. 4. The state ( $|\psi\rangle = |a, b\rangle$ ) transformation for this gate can be represented as follows.

$$|\psi\rangle \rightarrow U_S|\psi\rangle = U_S|a, b\rangle = |b, a\rangle.$$

Or in other words,

$$|a, b\rangle \rightarrow |a, a \oplus b\rangle \\ \rightarrow |a \oplus (a \oplus b), a \oplus b\rangle = |b, a \oplus b\rangle \\ \rightarrow |b, (a \oplus b) \oplus b\rangle = |b, a\rangle$$

Similar to CNOT gate, a SWAP gate can have also control Qbits. When the number of control Qbit is one, the gate

(CSWAP) becomes the well known Fredkin gate. The Fredkin gate is not only a reversible gate but also conservative, i.e. it is universal as well. Similar to Toffoli gate, the Fredkin gate also operates in space  $H_3$ . The swapping operation between the two target Qbits is performed when the control Qbit is  $|1\rangle$  (active high) or  $|0\rangle$  (active low). This means that the circuit can have two configurations: one is active high and other is the active low. The operational matrices (unitary matrices of dimension  $2^3 \times 2^3$ ) of these two circuit configurations are different. The circuit configurations along with their unitary matrices for active high and active low are shown in Fig. 5 and Fig. 6 respectively.

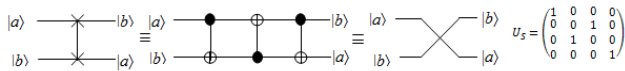


Fig. 4 Left to right: SWAP gate, equivalent circuit, equivalent symbol, SWAP gate matrix.

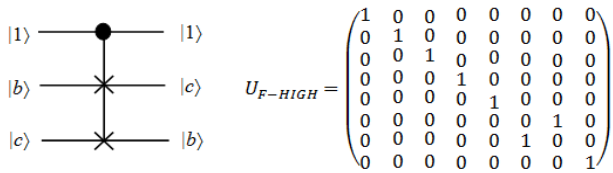


Fig. 5 Active high configuration of Fredkin gate and its matrix.

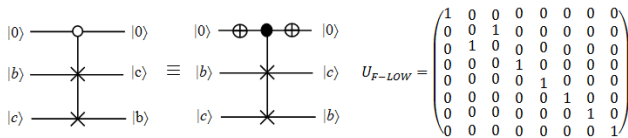


Fig. 6 Active low configuration of Fredkin gate, its equivalent circuit and its matrix.

Thus, when the number of control Qbit is increased to two, in a SWAP gate (i.e. for  $C^2$ SWAP gate), there exists four configurations for the swapping to be performed between the two target Qbits. The four circuit configurations along with their matrices are shown in Fig. 7.

### 3. Synthesis of QMUX

A CMUX consists of more than one input and only one output. The inputs of the multiplexer are two types: select inputs and data inputs. Depending upon the select inputs, at a time, only one of the data inputs is selected and sent to the output. If there are  $d$  data inputs in the circuit, then one needs at least  $s$  number of select input such that,  $2^s \geq d$ . This is the reason i.e. why commonly  $2^n$  (where,  $n = 1, 2, \dots$  etc.) number of data inputs are considered in the multiplexer circuit design. Unlike CMUX, the number of

outputs of a QMUX is equal to the total number of inputs (which is valid to any quantum circuit) of the circuit. Among the  $(s+d)$  number of outputs, only one output shows the desired multiplexing property. For optimization we have designed the QMUX in such a way that the multiplexing output will be available at  $D_{00}$ .

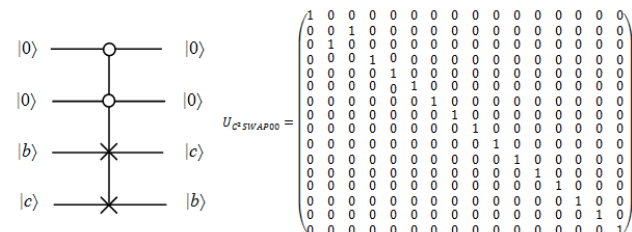
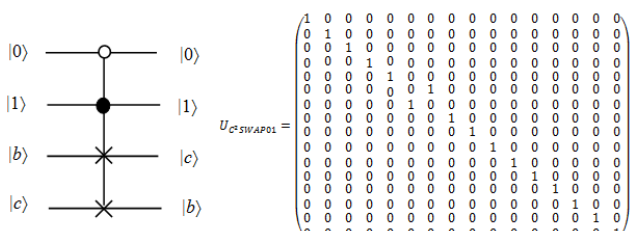
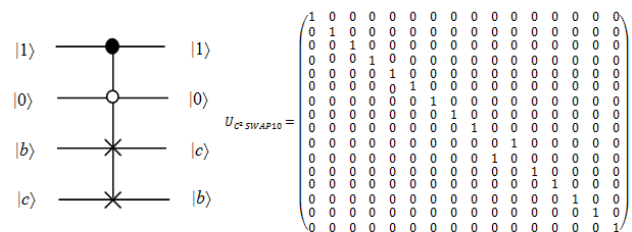
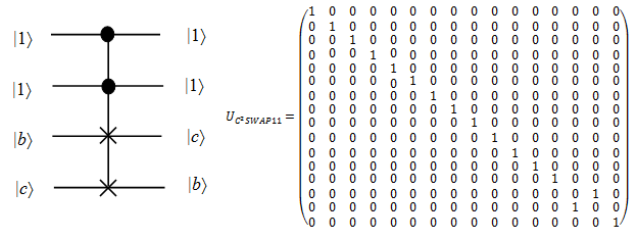


Fig. 7 Four configurations of the  $C^2$ SWAP gate and their matrices. Circuit swaps the target Qbits when both the control Qbits are set to  $|1\rangle$ ,  $|1\rangle$  for the configuration (a);  $|1\rangle$ ,  $|0\rangle$  for the configuration (b),  $|0\rangle$ ,  $|1\rangle$  for the configuration (c) and  $|0\rangle$ ,  $|0\rangle$  for the configuration (d).

The matrix (M) of a QMUX is block diagonal [4]. If the QMUX consists of  $s$  and  $d$  number of select inputs and data Qbits respectively, then the matrix M will be a block diagonal matrix having  $2^s$  blocks, each of size  $2^d \times 2^d$  [4]. Hence, the dimension of the matrix M will be  $2^{d+s} \times 2^{d+s}$ . A typical matrix for QMUX is shown below.

$$U_{MUX} = \begin{pmatrix} U_0 & \dots & \\ \vdots & \ddots & \vdots \\ & \dots & U_{n-1} \end{pmatrix}; \text{ where, } n-1 = 2^s.$$

### 3.1 Synthesis of 2:1 QMUX Circuit

The function of a 2:1 QMUX circuit can be performed by a Fredkin gate. A Fredkin gate can have two possible configurations (as shown earlier). Thus the possible two configurations along with their matrices of 2:1 QMUX can be found in Fig. 5 and Fig. 6. In these 2:1 QMUX circuits, the top most input Qbit is the control Qbit (or select input) and rest of two inputs are the data Qbits. Among the three output Qbits, the middle Qbit is the multiplexed Qbit. Let the states of the input Qbits for the circuit shown in Fig. 5 are:  $|S_0\rangle$ ,  $|D_{i0}\rangle$  and  $|D_{i1}\rangle$  (from top to bottom) and the same for the output Qbits are:  $|S_0\rangle$ ,  $|D_{i0}\rangle$  and  $|D_{i1}\rangle$  respectively, then we can express the state of the multiplexed output as:  $|D_{i0}\rangle = |D_{i0} \oplus S_0 (D_{i1} \oplus S_0 D_{i0})\rangle$ .

Let us consider the matrix  $U_{F-HIGH}$  in Fig. 5 to explain its block diagonal nature. In this case, the number of select Qbit,  $s=1$  and number of data input,  $d=2$ . Hence the number of the blocks in the said matrix is  $2^s = 2$ . The size of each block is  $2^d \times 2^d = 4 \times 4$ . To show the blocks of the matrix  $U_{F-HIGH}$ , it is rewritten as:

$$U_{F-HIGH} = \begin{matrix} & \begin{matrix} S_0=0 & & & & S_0=1 & & & & \end{matrix} \\ \begin{matrix} U_p \\ \vdots \\ U_1 \end{matrix} & \begin{pmatrix} 000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{matrix}$$

In the above expression, one can see the two blocks:  $U_0$  and  $U_1$  corresponds to the single select input  $S_0 = 0$  or  $S_0 = 1$ . Note that the size of the each block is  $4 \times 4$ . In a more simplified way, the above matrix can be written as:

$$U_{F-HIGH} = \begin{pmatrix} U_0 & O \\ O & U_1 \end{pmatrix}$$

### 3.2 Synthesis of 4:1 QMUX Circuit

This circuit consists of 6 inputs among which 2 inputs are select inputs and 4 inputs are data inputs. The circuit of the 4:1 QMUX is shown in Fig. 8. The truth table of the circuit is shown in Table 1. In the truth table, each input Qbit is shown as 1 or 0 for simplicity, however in reality, each Qbit is a state vector either  $|0\rangle$  or  $|1\rangle$  (or superposition of these two states) as shown earlier.

The dimension of the vector space ( $\mathcal{H}_6$ ) which represents the 4:1 QMUX is  $2^6$ . So for this circuit

$$\begin{aligned} \mathcal{H}_6 &= \text{span}_{\mathbb{C}}(|b\rangle; b \text{ is bitstring of length } 6) \\ &= \text{span}_{\mathbb{C}}(|b_1\rangle, |b_2\rangle \dots |b_6\rangle) \\ |b_i\rangle &= |S_1 S_0 D_{i0} D_{i1} D_{i2} D_{i3}\rangle; \\ \text{Where, } S_1, S_0, D_{i0}, D_{i1}, D_{i2}, D_{i3} &\in \{0,1\}. \end{aligned}$$

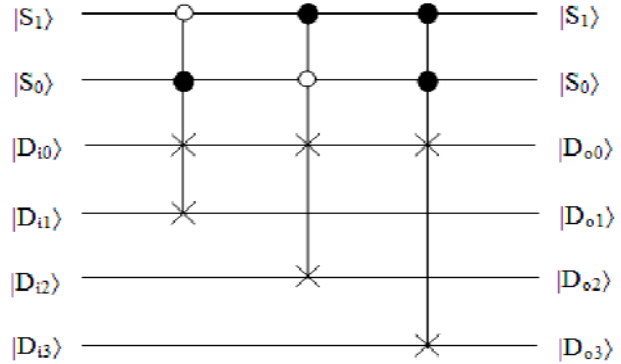


Fig. 8 Circuit of the 4:1 QMUX. Here,  $S_1, S_0$  are two select input Qbits and  $D_{i3}$  to  $D_{i0}$  are the four data input Qbits. The quantum wire  $D_{i0}$  shows the multiplexing output.

Table 1: Truth table of 4:1 QMUX.

$S_1$	$S_0$	$D_{i3}$	$D_{i2}$	$D_{i1}$	$D_{i0}$	Output( $D_{i0}$ )
0	0	0	0	0	0	$D_{i0}$
0	0	0	0	0	0	$D_{i0}$
.	.	.	.	.	.	$D_{i0}$
.	.	.	.	.	.	$D_{i0}$
0	0	1	1	1	1	$D_{i0}$
0	1	0	0	0	0	$D_{i1}$
0	1	0	0	0	1	$D_{i1}$
.	.	.	.	.	.	$D_{i1}$
.	.	.	.	.	.	$D_{i1}$
0	1	1	1	1	1	$D_{i1}$
1	0	0	0	0	0	$D_{i2}$
1	0	0	0	0	1	$D_{i2}$
.	.	.	.	.	.	$D_{i2}$
.	.	.	.	.	.	$D_{i2}$
1	0	1	1	1	1	$D_{i2}$
1	1	0	0	0	0	$D_{i3}$
1	1	0	0	0	1	$D_{i3}$
.	.	.	.	.	.	$D_{i3}$
.	.	.	.	.	.	$D_{i3}$
1	1	1	1	1	1	$D_{i3}$

Now, if  $|M\rangle$  is a state of the QMUX, then

$$|M\rangle \rightarrow \sum_{b \in \mathbb{B}^6} \alpha_i |b\rangle \rightarrow \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{2^6} \end{pmatrix};$$



Where,  $|\alpha_i|^2$  is the probability of the state of  $i^{\text{th}}$  element. So, a state of this space is basically realized by the above column vector. Similar to the previous multiplexer circuit, the matrix  $U_{4:1 \text{ QMUX}}$  in the block diagonal form can be written as:

$$U_{4:1 \text{ QMUX}} = \begin{pmatrix} U_0 & & & \\ & U_1 & & \\ & & U_2 & \\ & & & U_3 \end{pmatrix}$$

Where  $U_0, U_1, U_2$  and  $U_3$  are the four diagonal blocks, each of size  $2^4 \times 2^4$ . So here  $U_0, U_1, U_2$  and  $U_3$  are operated on data inputs  $D_{i0}, D_{i1}, D_{i2}$  and  $D_{i3}$  when  $S_0=0, S_1=0; S_0=0, S_1=1; S_0=1, S_1=0; S_0=1, S_1=1$  respectively.

In order to understand the operation of the QMUX in detail, the circuit is decomposed into nine functional blocks using LNN method. The LNN is often considered as an appropriate technique to scalable quantum architecture [6]. The decomposition of the circuit (using LNN method) is shown in Fig. 9.

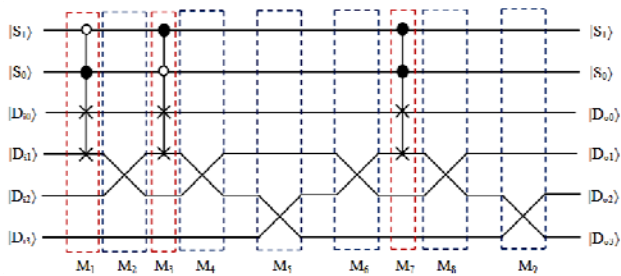


Fig. 9 Decomposition of 4:1 QMUX circuit into nine functional blocks.

Denoting  $I$  as  $2 \times 2$  identity matrix, the matrices of the nine blocks are:

$$M_1 = U_1 \otimes I^{\otimes 2} \text{ (where, } U_1 = U_C^2 \text{swap}_{01}, \text{ see Fig. 7c)}$$

$$M_2 = I^{\otimes 3} \otimes U_S \otimes I \text{ (for } U_S \text{ see Fig. 4)}$$

$$M_3 = U_2 \otimes I^{\otimes 2} \text{ (where, } U_2 = U_C^2 \text{swap}_{10}, \text{ see Fig. 7b)}$$

$$M_4 = M_2$$

$$M_5 = I^{\otimes 4} \otimes U_S$$

$$M_6 = M_2$$

$$M_7 = U_3 \otimes I^{\otimes 2} \text{ (where, } U_3 = U_C^2 \text{swap}_{11}, \text{ see Fig. 7a)}$$

$$M_8 = M_2$$

$$M_9 = M_5$$

The operation of each block is represented by a  $2^6 \times 2^6$  matrix. The resultant matrix  $U_{4:1 \text{ QMUX}}$  of the 4:1 multiplexer circuit is, therefore the multiplication of the nine matrices, i.e.

$$U_{4:1 \text{ QMUX}} = M_9 \times M_8 \times M_7 \times M_6 \times M_5 \times M_4 \times M_3 \times M_2 \times M_1$$

The input and output of the 4:1 QMUX circuit is a column of matrix of element 64 ( $2^{s+d} = 2^6 = 64$ ). If  $\Gamma$  represents an output matrix corresponding to the input matrix  $\Lambda$  of the circuit, then one can write,  $U_{4:1 \text{ QMUX}} \Lambda = \Gamma$ . This expression can also be written in the following form:

$$U_{4:1 \text{ QMUX}} \cdot \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_r \\ \vdots \\ \alpha_{r+f} \\ \vdots \\ \alpha_{2^n} \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{r+f} \\ \vdots \\ \alpha_r \\ \vdots \\ \alpha_{2^n} \end{pmatrix}; r \geq 17 \quad (4)$$

In the Equation (4),  $n = s+d = 6$  for 4:1 QMUX and the constraint,  $r \geq (2^d + 1)$  or  $r \geq 17$  is obtained as follows.

In Equation (4), the 17<sup>th</sup> element is the square-root of the probability of the input state  $|010000\rangle$ . The 17<sup>th</sup> position represents an element which changes its position according to combinations of select inputs other than all 0s. Here  $|\alpha_r|^2$  stands for probability of  $r^{\text{th}}$  state which changes according to the circuit operations and swaps with  $(r+f)^{\text{th}}$  state.

In order to check the reversibility of circuit, the unitary property of the matrix  $U_{4:1 \text{ QMUX}}$  is checked. It is found that the relation:  $(U_{4:1 \text{ QMUX}}) \cdot (U_{4:1 \text{ QMUX}})^T = I$  is valid for the circuit and hence the proposed multiplexer circuit is a reversible circuit.

### 3.3 Higher Order QMUX Synthesis

Considering the 4:1 QMUX circuit as reference, it is not difficult to construct a higher order multiplexer. By looking the sequences of 'O' and '●' and '×' in Fig. 8, one can easily construct an 11-input 8:1 QMUX circuit. Such an 8:1 QMUX circuit is shown in Fig. 10.

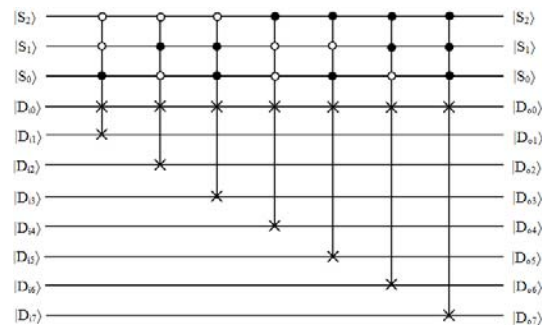


Fig. 10 An 11-input optimized 8:1 QMUX circuit.

Hence, it is possible to generalize the QMUX circuit for an  $n:1$  QMUX, where  $n = 2^r$ ,  $r$  is the number of select inputs. Such an optimistic generalized circuit is shown in Fig. 11.

### 4. Quantum Cost

Quantum cost is a measure of efficiency of a quantum circuit and it is commonly expressed in terms of Qgate. Using the LNN method, the quantum cost of different quantum gates is given in Table 2 [7].

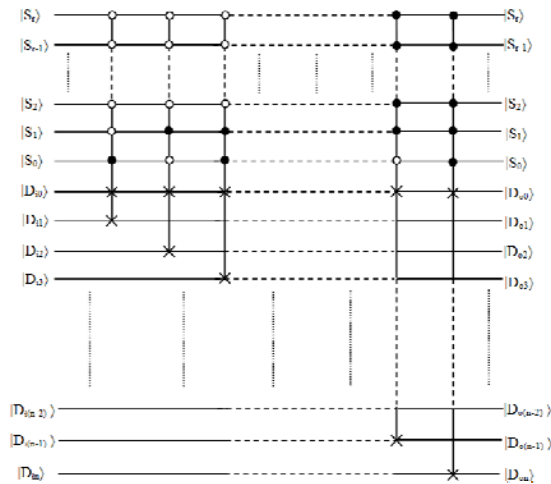


Fig. 11 Optimized n:1 QMUX circuit.

Table 2: Quantum cost of different Qgate [7].

Name of the Gate	Quantum cost
Controlled NOT	1
Controlled V	1
2-input SWAP	3
3-input Toffoli	9
4-input Toffoli	27
5-input Toffoli	45
3-input Fredkin	11
4-input Fredkin	29
5-input Fredkin	47

Similar technique is used to calculate the quantum cost for QMUX circuits and the summary of the cost calculation is presented in Table 3.

Table 3: Quantum cost for the QMUX circuits.

Order of Mux	No. of Select /control inputs(S)	No. of Control swap Gates(G)	Total No. of input lines (T)	No. of SWAP Gates according to each block (No. of blocks B=(n-1))											Total No. swap of Gates	Total No. of Gates (G+T <sub>swap</sub> )	Quantum Cost of QMUX
				B <sub>1</sub>	B <sub>2</sub>	B <sub>3</sub>	B <sub>4</sub>	B <sub>5</sub>	B <sub>6</sub>	B <sub>7</sub>	B <sub>8</sub>	...	B <sub>n-1</sub>				
2:1	1	1	3	0											0	1	11
4:1	2	3	6	0	2	4									6	9	109
8:1	3	7	11	0	2	4	6	8	10	12					42	49	473
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

n:1	$\frac{\ln(\pi)}{\ln 2}$	n-1	S+G+1	B <sub>N1</sub>	B <sub>N2</sub>	B <sub>N3</sub>	B <sub>N4</sub>	B <sub>N5</sub>	B <sub>N6</sub>	B <sub>N7</sub>	B <sub>N8</sub>	...	B <sub>N(n-1)</sub>	T <sub>nswap</sub>	G <sup>2</sup> =(n-1) <sup>2</sup>	X
-----	--------------------------	-----	-------	-----------------	-----------------	-----------------	-----------------	-----------------	-----------------	-----------------	-----------------	-----	---------------------	--------------------	------------------------------------	---

Where,

$B_{N1} = 0$

$B_{N2} = 2 \times 1; B_{N3} = 2 \times 2; \dots; B_{N(n-1)} = 2 \times (n-2)$

$T_{nswap} = [(n-2) \times (2 + B_{N(n-1)})] / 2 \Rightarrow T_{nswap} = G^2 - G$

Cost X =  $2S(2^{(S-1)} - 1) + (18S - 7)(2^S - 1) + 3 \times$

$T_{nswap} = R + (18S - 7)(2^S - 1) + 3 \times T_{nswap}$

Where, R =  $2S(2^{(S-1)} - 1)$  is the no. of QNOT gates required for a C<sup>n</sup>SWAP gate.

### 5. Conclusion

We have synthesized optimum QMUX circuit. In order to construct the quantum multiplexer, some physically realizable quantum gates are used and ‘Octave’ programming tool is used to present the functionality of the circuits. The matrix formulations and operational behavior of the circuits are presented in details. Our procedure shows the ability to construct a general n:1 QMUX.

### References

- [1] D. Mukhopadhyay and A. Si, “Quantum multiplexer design and optimization applying genetic algorithm” International Journal of Computer Science Issues, Vol. 7, Issues 5, 2010, pp. 360-366.
- [2] Mozammel H. A. Khan, “Design of reversible/quantum ternary multiplexer and demultiplexer”, Engineering Letters, Vol. 13:2, 2006, pp. EL\_13\_2\_3.
- [3] Vivek V. Shende, Aditya K. Prasad, Igor L. Markov and John P. Hayes, “Synthesis of reversible logic circuits”, IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems, Vol. 22, No. 6, 2003, pp. 710-722.
- [4] Vivek V. Shende, Stephen S. Bullock and Igor L. Markov, “Synthesis of quantum logic circuits”, IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems, Vol. 25, No. 6, 2006, pp. 1000-1010.
- [5] K. N. Patel, I. L. Markov and J. P. Hayes, “Optimal synthesis of linear reversible circuits”, Quantum Information and Computation, Vol. 8, No. 3&4, 2008, pp. 0282-0294.
- [6] B. Kane, “A silicon-based nuclear spin quantum computer”. Nature, Vol. 393, 1998, pp. 133-137.
- [7] Marek Perkowski, Martin Lukac, Dipal Shah, and Michitaka Kameyama; “Synthesis of quantum circuits in Linear Nearest Neighbor model using Positive Davio Lattices” Facta Uni. Ser. Elec. Energ., Vol. 24, No. 1, 2011, pp. 73-89.

**First Author** Dr. Arijit Roy has attended many prestigious institutes like IITs and Infineon Technologies in his career. He obtained the Ph.D. degree from School of Electrical and Electronic

Engineering, Nanyang Technological University, Singapore. His publication includes regular paper, review article, conference article etc. in international journals and conferences. His paper in Symposium On Electronics – 2004 (Singapore) won the second best paper award. Among his various publications, a 75-journal-page-size review article on “Electromigration” is published in a journal of impact factor 17.731. So far his research works have been cited more than one hundred times (excluding self citations) and his present research h-index is five. In 2011, Dr. Roy authored a research monograph, titled “Electromigration in Cu Interconnects, The Driving Force Formalism: Modeling and Experiments”, published by Lambert Academic Publishing, Germany (ISBN: 978-3-8454-1292-4). Dr. Roy is also involved as reviewer, advisory board member etc. for many international publishers. Dr. Roy is frequently invited for invited/plenary lectures by many organizations. His research interests include Micro- and Nano-electronics, Microelectronic Reliability, Quantum Circuits and Single Electron Transistor. Presently he is Assistant Professor and Head of The Electronics Department at West Bengal State University (Barasat, India).

**Second Author** Mr. Dibyendu Chatterjee completed his B.Sc. degree in Electronics (Hons.) and stood First Class First (gold medalist) from Bankura Christian College (under Burdwan University, West Bengal, India) in 2006 and M.Sc. degree in Electronic Science from Jadavpur University (West Bengal, Kolkata, India) in 2008. He is also qualified prestigious examinations such as NET (National Eligibility Test, conducted by UGC govt. of India) and GATE (Graduate Aptitude Test in Engineering, conducted by IITs) in Electronics. Presently Mr. Chatterjee is pursuing Ph.D. at the Department of Electronics, West Bengal State University (Barasat, India). His main area of research is Quantum Computation.

**Third Author** Mr. Subhasis Pal is graduated as a student of Electronics (Hons.) from Vidyasagar University (West Bengal, India) in 2006 and M.Sc in Electronic Science from Jadavpur University (West Bengal, India) in 2008. He got first class first position (gold medalist) in graduation. Presently, he is pursuing Ph.D. at the Department of Electronics, West Bengal State University (Barasat, India). His area of research is Quantum Computation.

# Multi-Objective Evolutionary Computation Solution for Chocolate Production System Using Pareto Method

Alaa Sheta<sup>1</sup>, Abdel karim Baareh<sup>2</sup>, Mohamed Ababna<sup>3</sup>, Noor Khrisat<sup>4</sup>

<sup>1</sup>Computer Science Department, The World Islamic Science and Education (WISE) University  
Amman 11947, Jordan

<sup>2</sup>Computer Science Department, Ajloun College, Al-Balqa Applied University  
Ajloun 26816, Jordan

<sup>3</sup>Information Technology Department, Al-Balqa Applied University  
Salt 19117, Jordan

<sup>4</sup>Information Technology Department, Al-Balqa Applied University  
Salt 19117, Jordan

## Abstract

Solving manufacturing engineering problems normally involves variety of challenges. It is important to maximize profit, improve quality of a product mean while reduce losses and cost. This trade-off plays a vital role in solving many manufacturing optimization problem. The Chocoman Inc, USA produces varieties of chocolate bars, candy and wafer by means of raw materials. The objective of the company is to minimize its cost while maximizing the production of eight products. The formulation of this problem resulted in five functions to be optimized based twenty nine constraints to be satisfied. This is a typical Multi-objective Optimization Problems (MOPs). Many methods attempted to solve this problem. In this paper, we provide a comparison between the Scalarization and Pareto methods based Genetic Algorithms (GAs) to solve the chocolate production problem. GAs provides an outstanding solution.

**Keywords:** Multi-objective Optimization Problems (MOPs), Evolutionary Computation, Chocolate Production System, Scalarization Method, Pareto Method, GEATbx, Matlab.

## 1. Introduction

Most manufacturing engineering problems involve multiple-objectives. For example, minimize cost, maximize performance, maximize quality, reduce defected products etc. These are difficult but practical problems which normally happen [1]. GA was successfully used to solve variety of problems in system design, optimization and control. Genetic algorithms (GAs) are adaptive search procedures which were first introduced by J. Holland at Michigan University, USA 1975, and extensively studied by K. De Jong, D. Goldberg and others. GA found to be a well-matched tool for this class of problems. GAs is a population based approach which can optimize a complex optimization function given a fitness function to evaluate the goodness of a solution. Two universal approaches to solve multiple-objective optimization were introduced in the literature [2]. The first approach is to combine each

individual objective functions into a single composite function [3]. Solving a single objective function problem is visible with many methods such as the utility theory, weighted sum method [4], etc. There were many disadvantages reported on using the given methods. For example, in the case of utility function, finding the accurate weight of the function is not an easy task. Besides, the values of the weight are to be optimized not always robust. It is most likely affected by little uncertainty. The second approach is to find an entire Pareto optimal solution set. A Pareto optimal set is defined as the set of solutions that are non-dominated with respect to each other. The basic concepts of non-dominated and Pareto optimal solutions is explained by the following example. Solution S is said to dominate solution Y if all components of S are at least as good as those of Y, with at least one better component, and S is non-dominated if it is not dominated by any solution. Pareto optimal solution usually provides a more practical solution to engineering problem since they are always a trade-off between key parameters of the problem. The Pareto set size is always a function of the number of objective functions.

In this paper the main motivation for using Evolutionary Algorithms (EA's) to solve multi-objective optimization problems is because EA's deal simultaneously with a set of possible solutions (called population) which allows us to find several members of the Pareto optimal set in a single run of the algorithm, instead of having to perform a series of separate runs as in the case of the traditional mathematical programming techniques. Additionally the EA's does not require problem specific knowledge to carry out a search [2]. Our goal is to solve the well-known chocolate production system problem as a multi-objective optimization problem using Genetic Algorithms. We plan to use the Genetic and Evolutionary Algorithm Toolbox with Matlab (GEATbx) [5] to solve the problem. A comparison between the Scalarization and Pareto methods will be provided.

## 2. Statement of the problem

The multi-objective optimization problem can be defined as follows [10]. Our objective is to find the vector  $\bar{x} = [\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n]^T$  which will satisfy the  $m$  inequality constraints:

$$g_i(\bar{x}) \geq 0 \quad i=1, \dots, m \quad (1)$$

The  $p$  equality constraints:

$$h_i(\bar{x}) = 0 \quad i=1, \dots, p \quad (2)$$

and optimize the vector function

$$f(\bar{x}) = [f_1(\bar{x}), f_2(\bar{x}), \dots, f_k(\bar{x})]^T \quad (3)$$

$\bar{x} = [x_1, x_2, \dots, x_n]^T$  is the vector of decision variables. In other words, we want the set of all numbers which satisfy Equations (1) and (2) using the particular set  $x_1^*, x_2^*, \dots, x_k^*$  which yields the optimum values of all the objective functions.

## 3. Multi-objective Optimization Problem

Genetic Algorithm uses computational models of natural evolutionary processes in developing computer based problem solving systems. Solutions are obtained using operations that simulate the evolution of individual structures through mechanism of reproductive variation and fitness based selection. Due to their success at searching complex non-linear spaces and their reported robustness in practical applications, these techniques are gaining popularity and have been used in a wide range of problem domain, one of which is the multi-objective problem [6, 7, 8, 9, 10, 11].

Different methods were used to explore the multi-objective optimization, such as the classic method for integrating several criteria scalarization, also called aggregation of objectives, and the Pareto method. Multiple Pareto-optimal solutions can be captured in the GA population in a single simulation run. A wide number of problems have been solved in various multi-objective optimization applications [12, 13, 14, 15, 16]. New and improved GA implementations studies were also investigated in [17, 18, 19, 20].

### 3.1 Scalarization method

Multi-objective optimization problems can be solved in numerous ways; a direct one is to combine them into a single scalar value (e.g., adding them together). This techniques are normally known as "aggregating functions", because they combine (or "aggregate") all the objectives of the problem into a single one.

An example of this technique is the fitness function that is used to solve the following problem:

$$J = \sum_{i=1}^k w_i f_i(x) \quad (4)$$

where  $w_i > 0$  are the weighting coefficients representing the relative importance of the  $k$  objective functions of our problem [2]. We usually assume that it has a value of 1 ( $\sum_{i=1}^k w_i = 1$ ).

Aggregating functions are a very common tool used to develop a direct implementation for the multi-objective problem were a single objective is used in fitness assignment, so a single objective GA can be used with minimum modifications. The drawback of this technique is that not all Pareto-optimal solutions can be investigated when the true Pareto front is non-convex. That's way, the multi-objective GAs based on the weighed sum approach have difficulty in finding solutions uniformly distributed over a non-convex trade-off surface [21]. When a multi-objective problem is solved by means of single-objective optimization, only a point solution is obtained. The advantage of obtaining several solutions of equal value relating to a target vector is lost. For that reason the user must decide to either use the simple weighted sum or the approximation of the Pareto-optimal solutions [22].

### 3.2 Pareto method

The MOPs is sometimes combined into a single objective so that traditional optimization and the mathematical programming methods can be used. Alternatively, a Pareto optimal set is found. This is usually achieved by using an evolutionary algorithm such as GA [23]. The definition for such a problem with more than one objective function (say,  $f_j, j = 1, \dots, M$  and  $M > 1$ ), with two solutions  $x_1$  and  $x_2$  can have one of two possibilities: one dominates the other or none dominates the other. A solution  $x_1$  is said to dominate the other solution  $x_2$ , if both the following conditions are true [24]:

1. The solution  $x_1$  is no worse (say the operator  $<$  denotes worse and  $>$  denotes better) than  $x_2$  in all objective, or  $f_j(x_1) > f_j(x_2) \quad \forall \quad j = 1, \dots, M$  objectives.
2. The solution  $x_1$  is strictly better than  $x_2$  in at least one objective, or  $f_j(x_1) > f_j(x_2)$  for at least one  $j \in \{1, 2, \dots, M\}$ .

If any of the above condition is violated, the solution  $x_1$  does not dominate the solution  $x_2$ . To solve the production problem with five objective functions using

Scalarization, and Pareto methods, we plan to use the GEATbx Toolbox based on Matlab [5].

#### 4. Chocolate production system

In this section, we provide a description for the famous production system chocolate problem for a chocolate exporting company. The data for this problem have been adopted from the data-bank of Chocoman Inc, USA [25]. The firm Chocoman, Inc. manufactures produced 8 different kinds of chocolate products since there are 8 raw materials to be mixed in different proportions and 9 processes (i.e. facilities) to be utilized. The problem can be presented as multi objective functions with 8 parameters to be optimized and 29 constrained that should be satisfied at the end of the evolutionary process that finds the optimal set of parameters. The objective of this problem is to maximize the five objective functions with eight variables. The decision variables for the chocolate problems are defined as:

- $x_1$  = milk chocolate of 250g to be produced
- $x_2$  = milk chocolate of 100g to be produced
- $x_3$  = crunchy chocolate of 250g to be produced
- $x_4$  = crunchy chocolate of 100g to be produced
- $x_5$  = chocolate with nuts of 250g to be produced
- $x_6$  = chocolate with nuts of 100g to be produced
- $x_7$  = chocolate candy to be produced
- $x_8$  = chocolate wafer to be produced

#### MAXIMIZATION - FIVE OBJECTIVE FUNCTIONS

- F1 **Revenue**  
 $F_1 = 375x_1 + 150x_2 + 400x_3 + 160x_4 + 420x_5 + 175x_6 + 400x_7 + 150x_8$
- F2 **Profit**  
 $F_2 = 0.25x_1 + 0.1x_2 + 0.25x_3 + 0.1x_4 + 0.25x_5 + 0.1x_6$
- F3 **Market Share for Chocolate Bars**  
 $F_3 = 0.25x_1 + 0.1x_2 + 0.25x_3 + 0.1x_4 + 0.25x_5 + 0.1x_6$
- F4 **Units produced**  
 $F_4 = x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8$
- F5 **Plant utilization**  
 $F_5 = 1.65x_1 + 0.9x_2 + 1.975x_3 + 1.03x_4 + 1.75x_5 + 0.94x_6 + 4.2x_7 + 1.006x_8$

Subject to the constraints:

- C1:  $x_1 \leq 0.6x_2$
- C2:  $x_3 \leq 0.6x_4$
- C3:  $x_5 \leq 0.6x_6$
- C4:  $400x_7 + 150x_8 \leq 56.25x_1 + 22.5x_2 + 60x_3 + 24x_4 + 63x_5 + 26.25x_6$
- C5: (cocoa usage)

$$87.5x_1 + 35x_2 + 75x_3 + 30x_4 + 50x_5 + 20x_6 + 70x_7 + 12x_8 \leq 100000$$

- C6: (milk usage)  
 $62.5x_1 + 25x_2 + 50x_3 + 20x_4 + 50x_5 + 20x_6 + 30x_7 + 12x_8 \leq 120000$
- C7: (nuts usage)  
 $0x_1 + 0x_2 + 37.5x_3 + 15x_4 + 75x_5 + 30x_6 + 0x_7 + 0x_8 \leq 60000$
- C8: (confectionery sugar usage)  
 $100x_1 + 40x_2 + 87.5x_3 + 35x_4 + 75x_5 + 30x_6 + 210x_7 + 24x_8 \leq 200000$
- C9: (flour usage)  
 $0x_1 + 0x_2 + 0x_3 + 0x_4 + 0x_5 + 0x_6 + 0x_7 + 72x_8 \leq 20000$
- C10: (aluminum foil usage)  
 $500x_1 + 0x_2 + 500x_3 + 0x_4 + 0x_5 + 0x_6 + 0x_7 + 250x_8 \leq 500000$
- C11: (paper usage)  
 $450x_1 + 0x_2 + 450x_3 + 0x_4 + 450x_5 + 0x_6 + 0x_7 + 0x_8 \leq 500000$
- C12: (plastic usage)  
 $60x_1 + 120x_2 + 60x_3 + 120x_4 + 60x_5 + 120x_6 + 1600x_7 + 250x_8 \leq 500000$
- C13: (cooking facility usage)  
 $0.5x_1 + 0.2x_2 + 0.425x_3 + 0.17x_4 + 0.35x_5 + 0.14x_6 + 0.6x_7 + 0.096x_8 \leq 1000$
- C14: (mixing facility usage)  
 $0x_1 + 0x_2 + 0.15x_3 + 0.06x_4 + 0.25x_5 + 0.10x_6 + 0x_7 + 0x_8 \leq 200$
- C15: (forming facility usage)  
 $0.75x_1 + 0.3x_2 + 0.75x_3 + 0.30x_4 + 0.75x_5 + 0.30x_6 + 0.90x_7 + 0.36x_8 \leq 1500$
- C16: (grinding facility usage)  
 $0x_1 + 0x_2 + 0.25x_3 + 0x_4 + 0x_5 + 0x_6 + 0x_7 + 0x_8 \leq 200$
- C17: (wafer making facility usage)  
 $0x_1 + 0x_2 + 0x_3 + 0x_4 + 0x_5 + 0x_6 + 0x_7 + 0.3x_8 \leq 100$
- C18: (cutting facility usage)  
 $0.5x_1 + 0.1x_2 + 0.1x_3 + 0.1x_4 + 0.1x_5 + 0.1x_6 + 0.2x_7 + 0x_8 \leq 400$
- C19: (packaging facility usage)  
 $0.25x_1 + 0x_2 + 0.25x_3 + 0x_4 + 0.25x_5 + 0x_6 + 0x_7 + 0.1x_8 \leq 400$
- C20: (packaging 2 facility usage)

$$0.05x_1 + 0.3x_2 + 0.05x_3 + 0.3x_4 + 0.05x_5 + 0.3x_6 + 2.50x_7 + 0.15x_8 \leq 1000$$

C21: (labor usage)  
 $0.3x_1 + 0.3x_2 + 0.05x_3 + 0.3x_4 + 0.3x_5 + 0.3x_6 + 2.50x_7 + 0.25x_8 \leq 1000$

C22: (demand for MC 250)  $x_1 \leq 500$

C23: (demand for MC 100)  $x_2 \leq 800$

C24: (demand for CC 250)  $x_3 \leq 400$

C25: (demand for CC 100)  $x_4 \leq 600$

C26: (demand for CN 250)  $x_5 \leq 300$

C27: (demand for CN 100)  $x_6 \leq 500$

C28: (demand for Candy)  $x_7 \leq 200$

C29: (demand for Wafer)  $x_8 \leq 400$

The parameters  $x_1, \dots, x_8$  must be nonnegative (i.e.  $x_1, \dots, x_8 \geq 0$ ). This problem has been solved in [30] using a newly developed Matlab toolbox called Hybrid Optimization Genetic Algorithms (HOGA). Comparatively; we solve the same problem by using different toolbox (i.e. GEATbx) [5] using two famous methods. They are the Scalarization and the Pareto methods.

## 5. Experimental Setup and Results

The Genetic and Evolutionary Algorithm Toolbox for use with Matlab (GEATbx) [5] contains a broad range of tools for solving real-world optimization problems. They not only cover pure optimization, but also the preparation of the problem to be solved. We can use the GEATbx as follows:

Creating m-file1

- Write the objective function to describe the problem
- Write the problem constraints

Parameters setting:

- Number Variable Default :8
- for Scalarization method, Number Objective Default: 1
- for Pareto method , the Number Objective Default: 5
- Variable Bound (Min): 0, 0, 0, 0, 0, 0, 0, 0
- Variable Bound (Max): 500, 800, 400, 600, 300, 500, 200, 400

- Save m-file1.
- Create m-file2.

Write the function which operates m-file1

Parameters setting:

- Population size: 20, 50, 100
- Termination Max Generations: 20
- Selection Mechanism: Stochastic Universal Sampling (SUS)
- Selection Pressure: 1.7
- Recombination Name: Recombination discrete
- Recombination Rate: 0.6
- Mutation mechanism: real value Mutation
- Generation Gap: 0.9
- Mutation Rate: 0.01
- Saving m-file2
- When using Scalarization method, we give an equal weight of 0.2 for each objective function.

### 5.1 Scalarization method Results

The firm Chocoman, Inc. manufactures 8 different kinds of chocolate products since there are 8 raw materials to be mixed in different proportions and 9 processes to be utilized. The objective of the company is to maximize its profit, which is, alternatively, equivalent to maximizing the gross contribution to the company in terms of US\$. Thus, we need to find the optimal product mix within a set of constraints in the technical, raw material and market consideration. The goal of this problem is to maximize the objective function subject to a given set of constraints. In [30], the goal was to maximize the objective function presented below subject to the same constraints:

$$Z = 180x_1 + 83x_2 + 153x_3 + 72x_4 + 130x_5 + 70x_6 + 208x_7 + 83x_8 - 0.18x_{12} - 0.16x_{22} - 0.15x_{32} - 0.14x_{42} - 0.13x_{52} - 0.14x_{62} - 0.12x_{72} - 0.17x_{82} \quad (5)$$

Depending on Scalarization method which converts the problem from multi-objective function to single objective function; this problem can be solved by running the GEATbx for 20 generations. The obtained results using Scalarization were compared with Sequential Quadratic Programming (SQP) presented in [30]. The computed values of the parameters  $x_1, \dots, x_8$  along with the optimal value of the objective function are presented in Table 1.

We note that the optimal value of the objective function using Scalarization method is 150440 which are better than the optimal value 147000 which was obtained by using SQP [3]. In Figure 1 we show the convergence process by showing the relationship between the generations and the optimal objective function value. It is shown that the function to be maximized reached the optimal value after 20 generations.

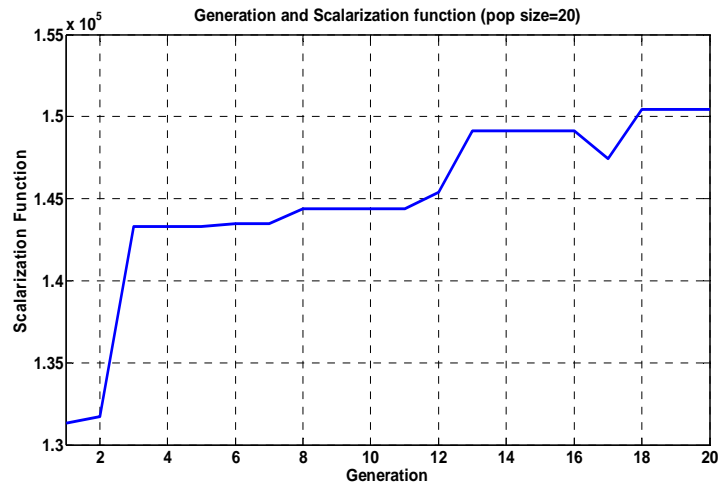


Fig 1. Optimal function value using Scalarization method

### 5.2 Pareto method results

The goal of this problem is to maximize the five objective functions that were presented previously depending on the Pareto method; this problem can be solved by running the GEATbx at different population sizes 20, 50, and 100. The sizes of the populations were selected arbitrary. In each case, we run GEATbx to find the optimal value of each function using various population sizes. The convergence process is shown in

each case. The obtained results for each function F1, ..., F5 is shown in Figures 2, 3, 4, 5 and 6. By the end of the evolutionary process, all curves convergence to the domain of the optimal solution. Although, the developed results with population size 100 looks the best. This gives us an indication which is increasing the population size might help in improving the performance of the developed results.

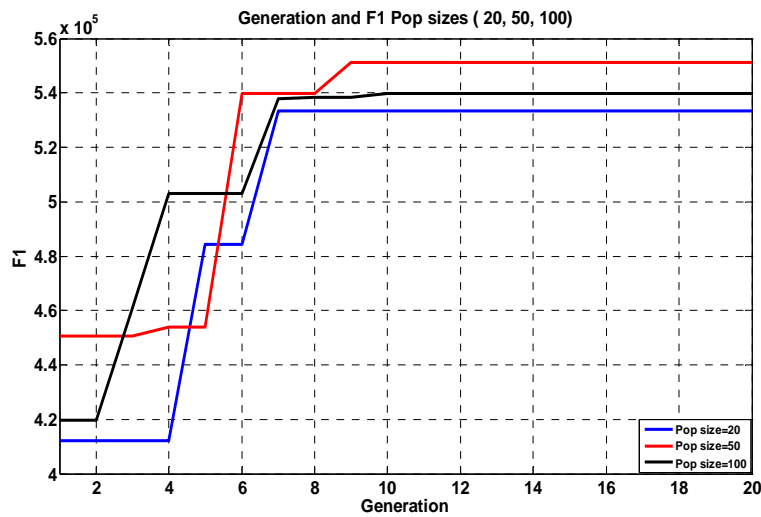


Fig 2. Optimal function value curve of F1 at different Pop Sizes



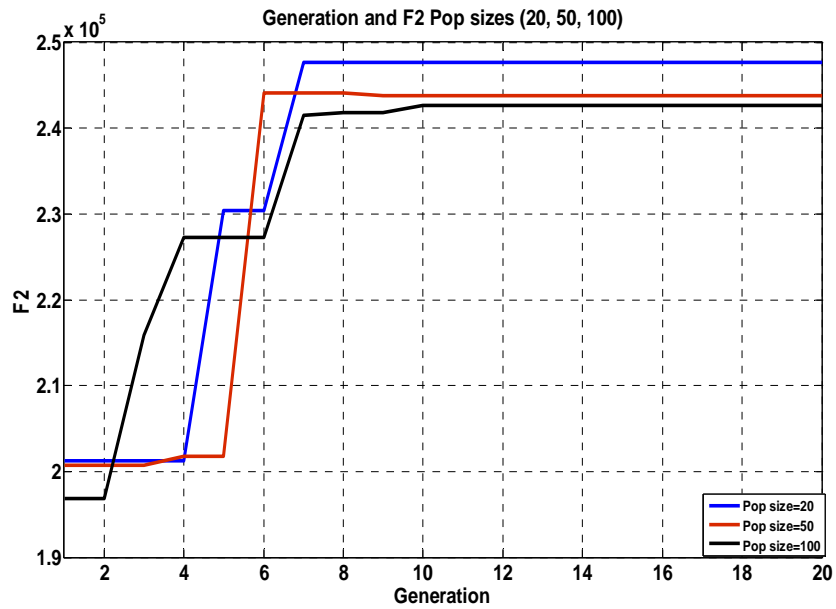


Fig 3. Optimal function value curve of F2 at different Pop Sizes

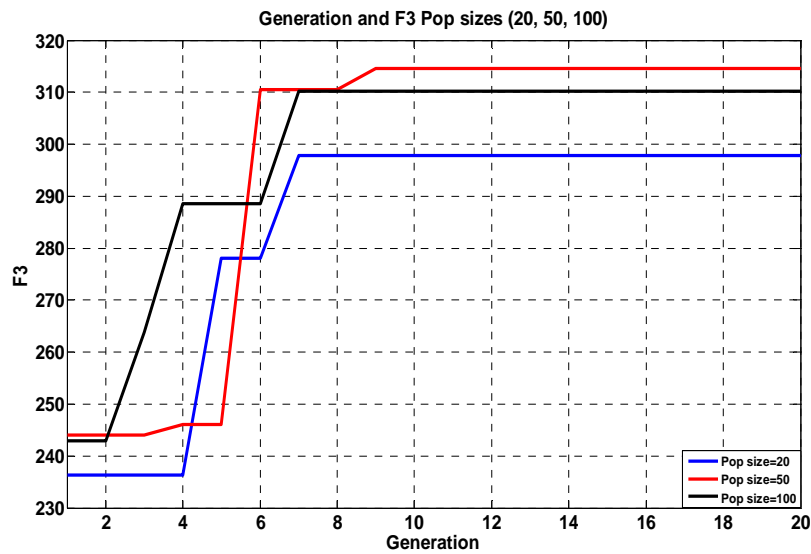


Fig 4. Optimal function value curve of F3 at different Pop Sizes

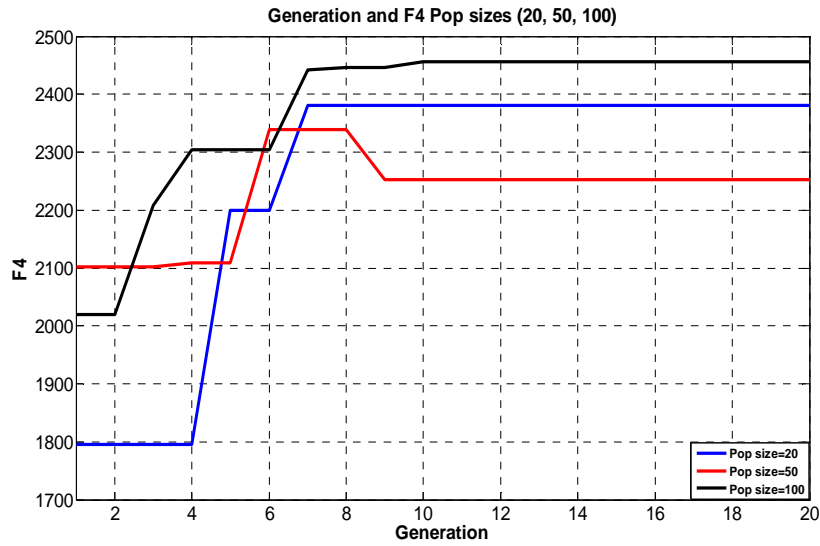


Fig 5. Optimal function value curve of F4 at different Pop Sizes

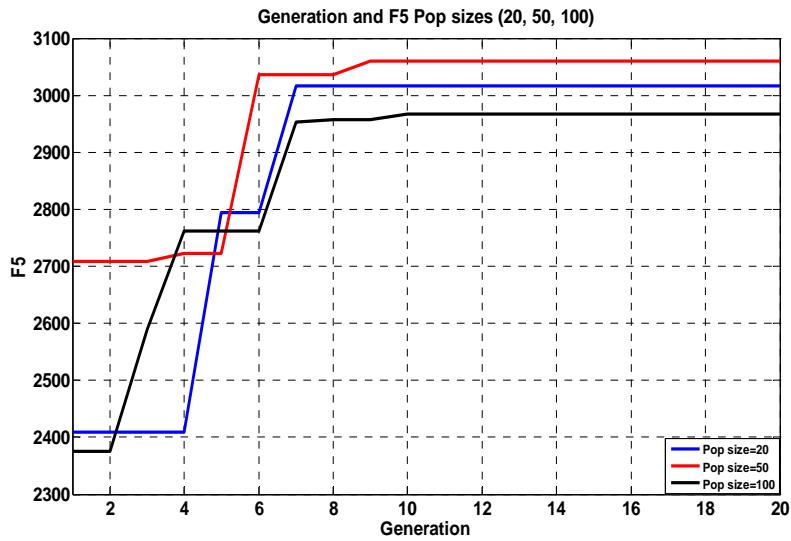


Fig 6. Optimal function value curve of F5 at different Pop Sizes

Table 1. Estimated values of the parameters for the Scalarization method and the Sequential Quadratic Programming (SQP) based GAs method [3]

Parameters	Scalarization Method	SQP-GAs method [30]
x1	122.65	217.55
x2	700.63	366.11
x3	46.564	246.34
x4	447.97	410.58
x5	279.59	226.05
x6	490.58	489.33
x7	134.83	76.781
x8	124.14	273.99
Optimal Value	150440	147000

Table2. Comparison between various Pop Sizes (20, 50, and 100)

	Pop Size =20	Pop Size =50	Pop Size =100
F1	533260	539680	539890
F2	247600	244070	242650
F3	297.84	310.51	310.24
F4	2381.2	2338.8	2456.4
F5	3015.7	3036	2967.8
Estimated states values			
x1	393.24	375.19	271.96
x2	739.63	640.27	681.61
x3	32.256	23.099	134.89
x4	328.29	529.72	520.2
x5	138.7	204.72	183.76
x6	500	427.54	424.11
x7	105.22	109.65	44.42
x8	143.85	28.568	195.4

## 6. Conclusion

In this paper, we provided a solution to the famous production system chocolate problem using both the Scalarization and Pareto methods. We compared our results with the results presented in [3]. Two methods were investigated to solve the production system problem. They are the Scalarization and Pareto methods. The developed results show an improvement in the produced optimal values to solve the MOP for the Chocolate production system than the recent reported results.

## References

[1] C. Coello, "An Empirical Study Of Evolutionary Techniques For Multiobjective Optimization In Engineering Design", PhD Dissertation, Department of Computer Science, Tulane University, 1996.  
 [2] C. Coello, Carlos A., Gary B. Lamont & David A. Van Veldhuizen, "Evolutionary Algorithms for Solving Multi-Objective Problems", Springer, New York, ISBN 978-0-387-33254-3, September 2007 (Second Edition).  
 [3] H. Turabieh, A. Sheta, and P. Vasant, "Hybrid Optimization Genetic Algorithm (HOGA) with Interactive Evolution to Solve Constraint Optimization Problems for Production Systems", in the International Journal of Computational Science, Vol. 1, No. 4, 2007, pp. 395-406.  
 [4] Y. Kim, O. Weck, "Adaptive Weighted Sum Method for Multiobjective Optimization", Proceeding of the 10th

AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, Albany, New York, 2004.  
 [5] H. Pohlheim, GEATbx Introduction Evolutionary Algorithms: Overview, Methods and Operators, (www.geatbx.com), November 2005.  
 [6] A. Lopez Jaimes and Carlos. A. C. Coello, "Multi-Objective Evolutionary Algorithms: A Review of the State-of-the-Art and some of their Applications in Chemical Engineering", in Rangaiah Gade Pandu (editor), Multi-Objective Optimization Techniques and Applications in Chemical Engineering, Chapter 3, pp. 61-90, World Scientific, Singapore, 2009.  
 [7] J. Horn, and N. Nafpliotis, and D. E. Goldberg, "A niched Pareto genetic algorithm for multiobjective optimization," Proceedings of the First IEEE Conference on Evolutionary Computation, 1994, Vol. 1, pp. 82-87.  
 [8] A. Lopez Jaimes and Carlos. A. C. Coello, "An Introduction to Multi-Objective Evolutionary Algorithms and some of Their Potential Uses in Biology", in Tomasz Smolinski, Mariofanna G. Milanova and Aboul-Ella Hassanien (editors), Applications of Computational Intelligence in Biology: Current Trends and Open Problems, pp. 79-102, Springer, Berlin, 2008.  
 [9] Fons, C. M., Fleming, P. J., Zitzler, E., Deb, K., and Thiele, L. (eds.): Evolutionary Multi-Criterion Optimization, Second International Conference (EMO) Lecture Notes in Computer Science, 2003, Vol. 2632, Berlin: Springer-Verlag.  
 [10] J. W. Eheart, S. E. Cieniawski, and S. Ranjithan, "Genetic-algorithm-based design of groundwater quality monitoring system," WRC Research Report No. 218. Urbana: Department of Civil Engineering, the University of Illinois at Urbana-Champaign, 1993.  
 [11] K. Mitra., K. Deb, and S. K. Gupta, "Multiobjective dynamic optimization of an industrial Nylon 6 semi-batch reactor using genetic algorithms", Journal of Applied Polymer Science ,Vol. 69, No. 1, 1998, pp. 69-87.  
 [12] G. T. Parks, and I. Miller, "Selective breeding in a multi-objective genetic algorithm", Proceedings of the Parallel Problem Solving from Nature, 1998, pp. 250-259.  
 [13] D. S. Weile, E. Michielssen, and D. E. Goldberg, "Genetic algorithm design of Pareto-optimal broad band microwave absorbers", IEEE Transactions on Electromagnetic Compatibility, 1996, Vol. 38, No. 4.  
 [14] C. M. Fonseca, and P. J Fleming, "Multi-objective optimization and multiple constraints handling with evolutionary algorithms - Part II: Application example", IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 1998, Vol. 28, No. 1, pp. 38-47.  
 [15] E. Mezura-Montes and Carlos. A. C. Coello, "Constrained Optimization via Multiobjective Evolutionary Algorithms", in Joshua Knowles, David Corne and Kalyanmoy Deb (Editors), Multi-Objective Problem Solving from Nature: From Concepts to Applications, pp. 53-75, Springer, ISBN 978-3-540-72963-1, 2008.  
 [16] E. Mezura-Montes, M. Reyes-Sierra and C. A. Coello, "Multi-Objective Optimization using Differential Evolution: A Survey of the State-of-the-Art", in Uday K. Chakraborty (editor), Advances in Differential Evolution, pp. 173-196, Springer-Verlag, Berlin, ISBN 978-3-540-68827-3.C, 2008.  
 [17] M. Fonseca, and P. J. Fleming, "An overview of evolutionary algorithms in multi-objective optimization", Evolutionary Computation, Vol. 3, No. 1, 1995, pp. 1-16.

- [18] C. Coello, Carlos A., C. Dhaenens and L. Jourdan (editors), "Advances in Multi-Objective Nature Inspired Computing", Springer, Berlin/Heidelberg, and ISBN: 978-3-642-11217-1, 2010.
- [19] A. Konak, A. Smith, and D. Coit, "Multi-Objective Optimization Using Genetic Algorithms: A Tutorial," Journal of Engineering Optimization, Vol. 36, No. 2, 2004, pp. 189-205.
- [20] T. Pulido, Gregorio and C. Coello, Carlos A. "A Constraint-Handling Mechanism for Particle Swarm Optimization", in 2004 Congress on Evolutionary Computation (CEC'), 2004, IEEE Portland Oregon June, Vol. 2, pp. 1396–1403.
- [21] C. Artemio, C. Coello, S. Dehuri and S. Ghosh (eds), "Swarm Intelligence for Multi-objective Problems in Data Mining", Springer, Berlin/Heidelberg, ISBN 978-3-642-03624-8, 2009.
- [22] S. Susanto, P .Vasant, A. Bhattacharya, and F. Pratikto, "Fuzzy LP with a nonlinear MF for product-mix solution: A case-based re-modelling and solution," in The 11th International Conference on Fuzzy Theory and Technology FTT in conjunction with 9th Joint Conference on Information Sciences, Kaohsiung, Taiwan, ROC, 2006, pp. 1305–1308.
- [23] A. Carlos, C. Coello, T. Pulido, Gregorio and M. Montes, Efr'en, "Current and Future Research Trends in Evolutionary Multiobjective Optimization", in Manuel Grana, Richard Duro, Alicia d'Anjou, and Paul P. Wang (editors), Information Processing with Evolutionary Algorithms: From Industrial Applications to Academic Speculations, pp. 213–231, Springer-Verlag, ISBN 1-8523-3866-0, 2005.
- [24] K. Deb, "Multi-Objective Genetic Algorithms: Problem Difficulties and Construction of Test Problems", Evolutionary Computation, Vol. 37, No. 3, 1999, pp. 205-230.
- [25] S. Susanto, P .Vasant, A. Bhattacharya, and F. Pratikto, "Fuzzy LP with a nonlinear MF for product-mix solution: A case-based re-modelling and solution", in The 11th International Conference on Fuzzy Theory and Technology FTT in conjunction with 9th Joint Conference on Information Sciences (JCIS ), Kaohsiung, Taiwan, ROC, 2006, pp. 1305–1308.



**Abdel Karim Baareh** received his B.Sc degree in Science from Mysore University India in 1992, completed a Post Graduate Diploma in Computer Application PGDCA from Mysore University India in 1993. He received his Master in Computer Application (MCA) from Bangalore University India in 1999. He received his Ph.D. in Informatics from Damascus University, Syria in 2008. Currently, Dr. Baareh is a faculty member with the Computer Science Department, Al-Balqa'a Applied University, Ajloun College, Jordan. He is the Chairman of the Applied Science Department at Ajloun College since 2009. His research interest includes Neural Networks, Fuzzy Logic, Image Processing and Genetic Algorithms.



**Mohammad F. Ababneh** received his Ph.D. in Computer Engineering from Cairo University, Egypt. Currently he is Associate Professor with the Computer Science Department and the Dean of the Faculty of Information Technology, Al-Balqa'a Applied University, Jordan. His research interest includes Artificial Intelligence, Computer Network, Virtual Reality, and Image Processing. Dr. Ababneh has been a program committee member for many national and international conferences. He is an active member for many national computer science societies.



**Alaa F. Sheta** received his B.E., M.Sc. degrees in Electronics and Communication Engineering from Faculty of Engineering, Cairo University in 1988 and 1994, respectively. He received his Ph.D. degree from the Computer Science Department, School of Information Technology and Engineering, George Mason University, Fairfax, VA, USA in 1997. Currently, Prof. Sheta is a faculty member with the Computer Science Department, The World Islamic Sciences and Education (WISE) University, Amman, Jordan. He is on leave from the Electronics Research Institute (ERI), Cairo, Egypt. He published over 80 papers, book chapters and two books in the area of image processing and evolutionary computations. He has been an invited speaker in number of national and international conferences. Prof. Sheta is a member of the IEEE Evolutionary Computations, ACM and ISAI societies. He is also the Vice President of the Arab Computer Society (ACS). His research interests include Evolutionary Computation, Modeling and Simulation of Dynamical Nonlinear Systems, Image Processing, Robotics, Swarm Intelligence, Automatic Control, Fuzzy Logic, Neural Networks and Software Reliability Modeling.

# Improve and Compact Population in XCSFCA using Polynomial Equation

Saeid Goodarzian<sup>1</sup>, Ali Hamzeh<sup>2</sup> and Sattar Hashemi<sup>3</sup>

<sup>1</sup>CE School, CSE and IT Department, Shiraz University, Shiraz, Iran

<sup>2</sup>CE School, CSE and IT Department, Shiraz University, Shiraz, Iran

<sup>3</sup>CE School, CSE and IT Department, Shiraz University, Shiraz, Iran

## Abstract

XCS is a rule-based evolutionary online learning system. XCSFCA is an extension of XCS where compute continuous actions directly from input states. In XCSFCA, computed actions of a classifier, demonstrated as a straight lines. But in very problems, the desired best action curves are not linear and there are arched; therefore a system with linear action computation needs a large population. This paper studies a new method for compute continuous actions directly from input states. In new proposed method action computes by polynomial equation. Consequently, each classifier represents a nonlinear action curve and the classifiers are more generalized. In comparison with XCSFCA, our method proves to be more efficient and smaller population size.

**Keywords:** XCSF, XCSFCA, continuous action, polynomial equation.

## 1. Introduction

Learning classifier systems define a new online model of genetic based machine learning where do not create a single solution while create multiple solutions, collected into a population, called *classifiers (rules)*. Classifiers adaptively change to new classifiers, to make more accurate decisions in return for inputs. The description in [4], describes three major machine learning problem types that learning classifier systems can solve them. In optimization problems, learning classifier system search problem landscape to find the best solution at hand. In classification problems, LCSs provide class labels to partition the given input patterns into different classes. Moreover, in reinforcement learning problems [17], LCSs propose an action for a situation and after the receipt of the reward, they apply reward back-propagation method to updated and improve classifiers.

Each *classifier cl* in learning classifier systems consists of a *Condition cl.C* part, an *Action cl.A* part and a *Reward Prediction cl.R* value. Classifier *cl* predicts reward *cl.R* given its condition *cl.C* is satisfied, and given further that action *cl.A* is executed.

Learning classifier system as an online method, in each time step, senses a situation of problem environment and then checks all classifiers in its population [P] to find which classifier condition can satisfy the current situation, called matching. All matched classifiers are inserted into a list called match set [M], and then according to their eligibility, one action would be selected and performed on the environment, the involved classifiers are inserted into another list called action set [A]. Regarding to the action effects, the payoff value is received and used to update the parameters of the classifiers in the action set [A].

XCS [20] is one of the most well-known Learning Classifier Systems which attracted many researchers nowadays [11]. At 2002, as a mail stone, Wilson has introduced one of the most promising extensions to XCS, called XCSF, in [23]. The most important modification in XCSF with respect to XCS is its ability to compute the environmental payoff using an approximation method instead of tuning a real number. The overall architecture of XCSF is based on XCS and is very similar to XCSI [22]; a version of XCS with continuous classifier condition. In XCSI, each classifier has interval condition instead of traditional binary ones.

Considering current researches in XCS realm, it can be said that condition and prediction parts are two basic components which are mostly been improved and investigated [5, 12, 13, 14, 15 and 21]. However, in recent researches, the action part becomes more popular due to its importance role that can directly affect very large application area such as classification, approximation, simulation, control etc. [7, 8, 9 and 16]. Briefly, it must be noted that almost in all proposed architecture for a learning classifier system, the action part consists of an integer indicating particular class or effects in the environment among a limited list of candidates. However, in a newly introduced classifier system called Generalized Classifier System [25], a novel representation for the action part, named the continuous actions, is introduced. This scheme is able to compute the action instead of selecting it and therefore extends the range of possible actions from a limited discrete set to a continuous range. To describe in

brief, it is worth mentioning that the structured of each classifier in GCS is formed as:

$$t(x, a) \Rightarrow p(x, a) \quad (1)$$

Where  $t(x, a)$  is the condition part of a particular rule and  $p(x, a)$  is its payoff prediction. In this paper, some improvements in XCSFCA will be investigated with the aim of increasing the performance and compacting the resulted population. Here, the classifier action is computed as a polynomial combination of the input and a vector of tuned coefficients.

The rest of this paper is organized as follows: in the next section we describe XCSF in brief, and then some relevant works on continuous action are reviewed. Then we described our proposed method and benchmark problems. At last, new method's results are presented and discussed.

## 2. XCSF in brief

XCSF [23] is a model of Learning Classifier System that extends the typical concepts of classifiers through the introduction of a computed classifier prediction. To develop XCSF, XCS has to be modified in three respects: (i) classifier conditions are extended for numerical inputs, as done in XCSI [22]; (ii) classifier are extended with a vector of weights  $\vec{w}$ , that are used to compute the classifier prediction; finally, (iii) The original update of the classifier prediction must be modified so that the weights are updated instead of the classifier prediction. These three modifications result in a version of XCS, XCSF [23] that maps numerical input into actions with an associated calculation prediction.

**Classifiers:** In XCSF, classifiers consist of a condition, an action and four main parameters. The condition specifies which input states the classifier matches; as in XCSI [22], it is represented by a concatenation of interval predicates,  $int_i = (l_i, u_i)$ , where  $l_i$  ("lower") and  $u_i$  ("upper") are integers, though they might be also real. The action specifies the action for which the payoff is predicted. The four parameters are: (i) The weight  $\vec{w}$ , used to compute the classifier prediction as a function of the current input; (ii) The prediction error  $\epsilon$ , that estimates the error affecting the classifier prediction; (iii) The fitness  $F$  that estimates the accuracy of the classifier prediction; (iv) The numerosity  $num$ , a counter used to represent different copies of the same classifier. The weight vector  $\vec{w}$  has one weight  $w_i$  for each possible input and an additional weight  $w_0$  corresponding to a constant input  $x_0$ , which is set as a parameter of XCSF.

**Performance Component:** XCSF works as XCS. At each time step  $t$ , XCSF builds a match set  $[M]$  containing the classifiers in the population  $[P]$  whose condition matches

the current sensory input  $s_t$ ; if  $[M]$  contains less than  $\theta_{mna}$  actions, covering takes place and creates a new classifier that matches the current inputs and has a random action. Each interval predicates  $int_i = (l_i, u_i)$  in the condition of a covering classifier is generated as  $l_i = s_t(i) - rand_1(r_0)$  and  $u_i = s_t(i) + rand_1(r_0)$ , where  $s_t(i)$  is the input value of state  $s_t$  matched by the interval  $[0, r_0]$  with  $r_0$  fixed integer. The weight vector  $\vec{w}$  of covering classifiers is initialized with zero values (note in the original paper, weight are initialized with values in  $[0, 1]$ ); all the other parameters are initialized as in XCS [20]. For each action  $a_i$  in  $[M]$ , XCSF computes the system prediction that estimates the payoff that XCSF expects when action  $a_i$  is performed. As in XCS, in XCSF the system prediction of action  $a$  is computed by the fitness weighted average of all matching classifiers that specify action  $a$ . However, in contrast with XCS, in XCSF the classifier prediction is computed as a function of the current state  $s_t$  and the classifier weight vector  $\vec{w}$ . Accordingly, in XCSF system prediction is a function of both the current state  $s$  and the action  $a$ . Following a notation similar to [6], the system prediction for action  $a$  in state  $s_t$ ,  $P(s_t; a)$ , is defined as equation 2:

$$P(s_t; a) = \frac{\sum_{cl \in [M]_a} cl.p(s_t) \times cl.F}{\sum_{cl \in [M]_a} cl.F} \quad (2)$$

Where  $cl$  is a classifier,  $[M]_a$  represents the subset of classifiers in  $[M]$  with action  $a$ ,  $cl.F$  is the fitness of  $cl$ ;  $cl.p(s_t)$  is the prediction of  $cl$  computed in the state  $s_t$ . In particular,  $cl.p(s_t)$  is computed as equation 3

$$cl.p(s_t) = cl.w_0 \times x_0 + \sum_{t>0} cl.w_i \times s_t(i) \quad (3)$$

Where  $cl.w_i$  is the weight  $w_i$  of  $cl$  and  $x_0$  is a constant input. The values of  $P(s_t; a)$  form the prediction array. Next, XCSF select s an action to perform. The classifiers in  $[M]$  that advocate the selected action are put in the current action set  $[A]$ ; the selected action in sent to the environment and a reward  $r$  is returned to the system together with next input state  $s_{t+1}$ .

**Reinforcement Component:** XCSF uses the incoming reward  $r$  to update the parameters of classifiers in action set  $[A]$ . First, the reward  $r$  is used to update the weight vector  $\vec{w}$  using a modified delta rule [19] as follows for each classifier  $cl \in [A]$ , each weight  $cl.w_i$  is adjusted by a quality  $\Delta w_i$  computed as equation (4):

$$\Delta w_i = \frac{\eta}{\|\vec{x}_{t-1}\|^2} (r - cl.p(s_{t-1})) x_{t-1}(i) \quad (4)$$

Where  $\eta$  is a correction rate and  $\vec{x}_{t-1}$  is defined as the input state vector  $s_{t-1}$  augmented by a constant  $x_0$  (i.e.  $x_{t-1} = \langle x_0, s_{t-1}(1), s_{t-1}(2), \dots, s_{t-1}(n) \rangle$ ) and

$\|\vec{x}_{t-1}\|^2$  is the norm of vector  $\vec{x}_{t-1}$  for further details refer to [23]. The values  $\Delta w_i$  are used to update the weights of classifier  $cl$  as equation 5.

$$cl.w_i \leftarrow cl.w_i + \Delta w_i \quad (5)$$

Then the prediction error  $\varepsilon$  is updated as equation 6:

$$cl.\varepsilon \leftarrow cl.\varepsilon + \beta(|r - cl.p(s_{t-1})| - cl.\varepsilon) \quad (6)$$

Where  $\beta$  is the learning rate. Classifier fitness is updated as in XCS. First, the raw accuracy  $\kappa$  of the classifiers in [A] is computed as equation 7.

$$cl.\kappa = \begin{cases} 1 & \text{if } cl.\varepsilon < \varepsilon_0 \\ \alpha \left(\frac{cl.\varepsilon}{\varepsilon_0}\right)^{-\nu} & \text{otherwise} \end{cases} \quad (7)$$

Where  $\varepsilon_0$  is a constant that controlled the acceptable values of prediction error  $\varepsilon$ . If  $cl.\varepsilon$  is less than  $\varepsilon_0$  the error is accepted and classifier is accurate ( $cl.\kappa = 1$ ), otherwise the accuracy of classifier  $cl$  is controlled by parameters  $\alpha$  and  $\nu$ . For represent efficient accuracy of each classifier,  $cl.\kappa'$  calculated respect to other classifier accuracies and repetition, so the raw accuracy  $\kappa$  is used to calculate the relative accuracy  $\kappa'$  as equation 8.

$$cl.\kappa' = \frac{cl.\kappa \times cl.uum}{\sum_{j \in [A]} cl_j.\kappa \times cl_j.num} \quad (8)$$

Finally, the relative accuracy  $\kappa'$  is used to update the classifier fitness as equation 9.

$$cl.F = cl.F + \beta(cl.\kappa' - cl.F) \quad (9)$$

An algorithmic description of the overall update procedure is reported in [13].

**Discovery Component:** The genetic algorithm in XCSF [23] works as in XCSI [22]. On a regular basis depending on the parameter  $\theta_{GA}$ , the genetic algorithm is applied to classifiers in [A]. It selects two classifiers with probability proportional to their fitness, copies them, and with probability  $\chi$  performs crossover on the copies; then, with probability  $\mu$  it mutates each allele. Crossover and mutation work as in XCSI [22]. The resulting offspring are inserted into the population, if the population size is exceeding the maximum size of population, deletion method performed to delete the excessive classifiers.

### 3. Related works on continuous actions in LCS

In LCSs, actions are typically fixed, discrete, and encoded by a set of symbols, e.g. {0, 1}, {"Left", "Right", and "Top", "Down"}, {"12", "15"... "63"}. However, continuous real-valued actions are desirable in many

applications especially where fine reactions are more important. This is very complicated for a LCS with discrete actions to handle the continuous real-valued action range; therefore, LCSs that can generate the actions according to sensory inputs are interested. This section reviews two notable investigations in this area.

#### 3.1 Generalized Classifier System (GCS)

In [25], Wilson described three distinct classifier system architectures for continuous action. Generalized classifier system (GCS) is more applicable and remarkable architecture in comparison with the other one. Many basic parts of GCS inherited from XCSF however extensions where permit continuous action describes here. Each classifier in GCS structured in format of:

$$t(x, a) \Rightarrow p(x, a) \quad (10)$$

Where  $t(x, a)$  is the condition part of a particular rule, and  $p(x, a)$  is its payoff prediction computed as a linear combination of weight vector  $\vec{w}$  and collected vector of input  $x$ , action  $a$  and a constant value named  $x_0$ . Satisfying  $t(x, a)$  is related to both values  $x$  and  $a$ . since for each time step  $t$  best action for each satisfied classifier is desirable, equation (11) was used to compute the best action:

$$a^*(x) = \max_{a \in A} P(x, a) | t(x, a) = true \quad (11)$$

Where  $a^*(x)$  is the best action for input  $x$  and  $A$  is a continuous range for valid actions.

In GCS, author inspired the idea from an investigation on XCSF where the condition parts of classifiers are represented as general hyper-ellipsoidal [5].

Satisfying  $t(x, a)$  is depends on the values of  $x$  and  $a$ . So, although to form the match set [M] the values of both  $x$  and  $a$  are needed, the action  $a$  is the system output and not available yet. So in exploration phase a random action  $a$  generated, but in exploitation phase a different methodology is used to find the  $a_{best}$  (the best action of a particular classifier) which is desirable for the learner. In exploitation phase, a classifier would be a member of [M] if its condition part matches the input  $x$  and has any value of  $a \in A$ . Suppose that  $a_l$  and  $a_u$  is the minimum and maximum value of  $a \in A$ . Since,  $p(x, a)$  is computed linearly. Thus, it is clear that through all possible values of  $a$  either  $a_l$  or  $a_u$  can lead to maximize  $p(x, a)$ , in other words one of  $p(x, a_l)$  or  $p(x, a_u)$  have the highest prediction value. So, either  $a_l$  or  $a_u$  will be selected as the  $a_{best}$ . Finally, for  $a^*$ , the system pick the best of all  $a_{best}$  that yield higher prediction among the others.

### 3.2 XCSF with computing continuous actions (XCSFCA)

XCSFCA is an extension of XCSF that can be applied on the environments where the action could be assumed as a computable function with respect to the environmental input. However, in the recent GCS [25], actions are selected from a continuous range; also in XCSFCA, actions are directly computed from a continuous function of the input. In XCSFCA, the classifier action  $cl.a$  is computed as equation (12) suggested:

$$cl.a(x, \zeta) = \zeta \cdot x'$$
$$x' = (x_0, x_1, \dots, x_n) \quad (12)$$

Where  $\zeta$  is a vector of action weights and  $x_0$  is a constant, also a vector of mutation rate  $cl.\sigma$  is added to each classifier to be used in action weight updating phase. An evolutionary strategy (ES [10]) evolves the action weights to compute actions that are more accurate. The XCSFCA principal changes as follow:

**The process of building the match set [M]:** XCSFCA builds a match set [M] containing the classifiers in the population [P] whose condition matches the current sensory input and its computed action  $cl.a$  belongs to the range of action  $a_{range}$ , where  $a_{range}$  is a range of acceptable values for actions.

**Covering operator:** classifier  $cl$  is accepted and inserted into population [P] and consequently into match set [M] if the computed action  $cl.a$  belongs to the allowed action range.

**Action selection in exploration:** the action with the highest prediction is selected.

## 4. Extension to XCSFCA by polynomial function

As described in section 3.2, XCSFCA uses a linear combination of action weight vector  $cl.\zeta$  and  $x'$  to compute actions in reply to specified environmental state  $x$ . The length of weight vector  $\zeta$  is related to problem landscape dimension. In XCSFCA classifiers action curve represented as straight line. Since the classifier action  $cl.A$  must align to desired action curve, while the problem is not complex and the relation between input  $x$  and best action  $a^*$  is uniformly ascending or descending, the linear actions is efficient. In more complex problems the curve of best action  $a^*$  is not uniformly ascending or descending and it is not linear like and it is arched. Therefore, XCSFCA must evolve a large population to handle the problem landscape entirely, because a particular classifier in XCSFCA just works finely in a small area of problem

landscape. If the classifiers activation area becomes undersized, more classifiers need to cover problem landscape. For these types of problems, in one side, producing a suitable population and in the other side evolving action weight vector  $cl.\zeta$  for a particular classifier is time consuming processes, also population size, proportionally grown up.

In this section, we describe that the action computation function can change to reach another mapping type of input  $x$  to action  $a$ .

We replaced action weight vector  $\zeta$  with action computation coefficients  $v = (v_0, v_1, v_2, \dots, v_n)$ . If our modification of XCSFCA, the classifier action  $cl.a$  is computed as a polynomial equation of  $v$  and the current input  $x$ . Degree of polynomial is related to problem, e.g. for a polynomial with degree 2 the classifier action computed as equation 13:

$$cl.a(x, v) = v_2 * x^2 + v_1 * x + v_0 \quad (13)$$

It is clear that, for a low dimensional problem, a polynomial function with opportune degree is more powerful to compute actions because the number of action computation coefficients  $v$  is related to polynomial degree and we can adjust the polynomial degree considering the problem complexity. In the other point of view, through using polynomial function for compute actions, a particular classifier is able to cover a larger subsection of problem landscape because this is more flexible and better to be aligned with best actions curve.

We expected that, modified XCSFCA using polynomial function can solve problems that are more complex, also it can compact the population size in the simplest problems. The promising results presented in this paper approve our claim.

## 5. Experimental Setup

This section described *frog1* and *frog2* problems from [24] and [18]. Also for better examination, our proposed method a new *frog* problem introduced. In comparison with *frog1* and *frog2*, the new *frog* problem is more difficult.

### 5.1 Frog problems

The *frog* problem introduced in [24] is a one-dimensional problem describing a frog who wants to catch a fly that is located at the distance  $d$ . The frog senses the distance  $d$  through a value named  $x$  calculated as equation (15), and jumps with respect to  $x$  then the frog receive the payoff using the function given as equation (14):



$$P(x, a) = \begin{cases} x + a & \text{if } x + a \leq 1 \\ 2 - (x + a) & \text{otherwise} \end{cases} \quad (14)$$

$$x(d) = 1 - d \quad (15)$$

As an extension to the original *frog* problem, the *frog2* problem is introduced in [12] where both payoff function and transformation of  $d$  to  $x$  are modified as equation (16) and (17):

$$P(x, a) = \begin{cases} xe^a & \text{if } a \leq -\ln x \\ x^{-1}e^{-a} & \text{otherwise} \end{cases} \quad (16)$$

$$x(d) = e^{-d} \quad (17)$$

Fig. 1 and Fig. 2 show the best action  $a^*$  of the *frog1* and *frog2* with respect to  $x$ . We introduce an extension of *frog* problem called *Frog3*, which is demonstrated in fig. 3(a, b). In *Frog3*, the payoff function is still continuous and nonlinear which is composed of two nonlinear forms, also the relation between  $x$  and  $a^*$  is not uniformly ascending or descending. The payoff function is computed as equation (18) and (19):

$$(x, a) = \begin{cases} \frac{a}{\frac{1}{2}(\sin(2\pi x)+1)} & \text{if } a \leq \frac{1}{2}(\sin(2\pi x)+1) \\ \frac{a - \frac{1}{2}(\sin(2\pi x)+1)}{\frac{1}{2}(\sin(2\pi x)+1) - 1} + 1 & \text{otherwise} \end{cases} \quad (18)$$

$$x(d) = \frac{1}{2}(\sin(2\pi d)+1) \quad (19)$$

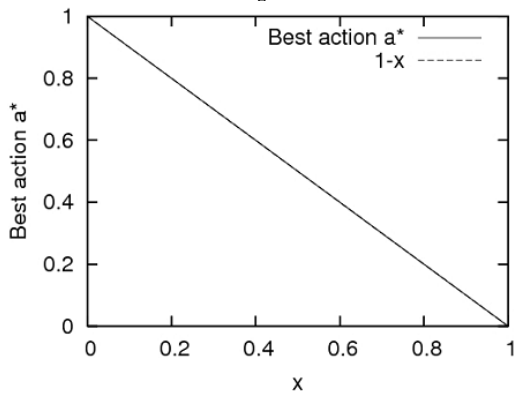


Figure 1: Best action  $a^*$  of *frog1*

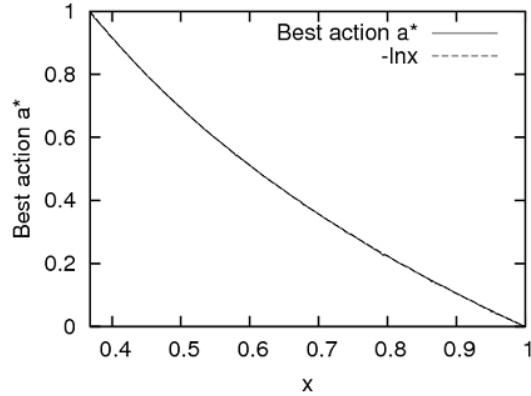


Figure 2: Best action  $a^*$  of *frog2*

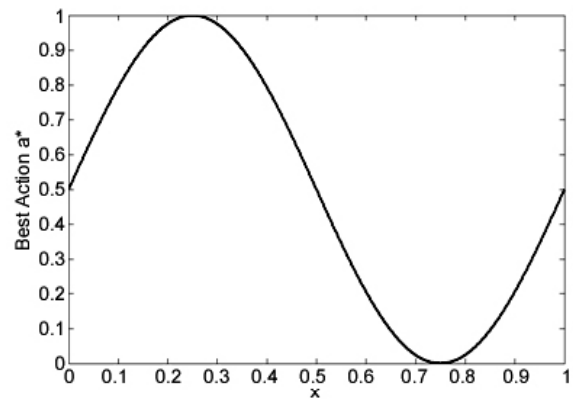


Figure 3(a): Best action  $a^*$  of *frog3*

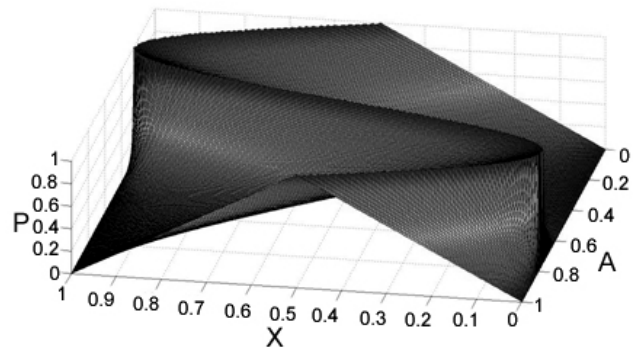


Figure 3(b): Payoff for all input  $x$  and action  $a$  of *frog3*

## 5.2 Simulation Setting

All the experiments discussed in this paper are performed following the standard design used in the literature [25]. To apply fair comparison of our results with other related work results, we used the minimal change to parameter setting. The parameters setting for the experiment were as follows:  $N = 2000$ ,  $\beta = 0.5$ ,  $\alpha = 0.1$ ,  $\eta = 0.2$ ,  $\delta = 0.1$ ,

$\theta_{GA} = 48$ ,  $\nu = 5$ ,  $\varepsilon_0 = 0.01$ ,  $\mu = 0.04$ ,  $\chi = 0.8$ ,  $\theta_{del} = 50$ . Both GA-subsumption and action set subsumption were not activated. Each run would stop after 100,000 explore problems. Explore and exploit problems are experimented one after one. As a polynomial with degree 2 is used to compute the actions, each classifier has an array with 3 elements  $v = (v_0, v_1, v_2)$  as action computation coefficients. The payoff would be received after applying the selected action. The fly positions were randomly selected from continuous range  $[0, 1]$ . The actions were computed within continuous range  $[0, 1]$ . To plot the best action  $a$ ,  $x$  was scanned from 0 to 1 and from  $e^{-1}$  to 1 and from 0 to 1 in *frog1*, *frog2* and *frog3* problems respectively, increased by 0.001. The best action curve is plotted, averaged over ten runs.

## 6. Results

### 6.1. Frog1's results.

Fig. 4 shows the system performance (in black), population size (in blue) and system error (in red). As it Fig.4 shows, the system performance is greater than 99% and system error drops to smaller than 1% after 21000 explore problems. The final population size of classifiers  $N$  is smaller than 19% of  $N$ , and it is worth to mentioning that the final population size is about half of XCSFCA population size showed in [18].

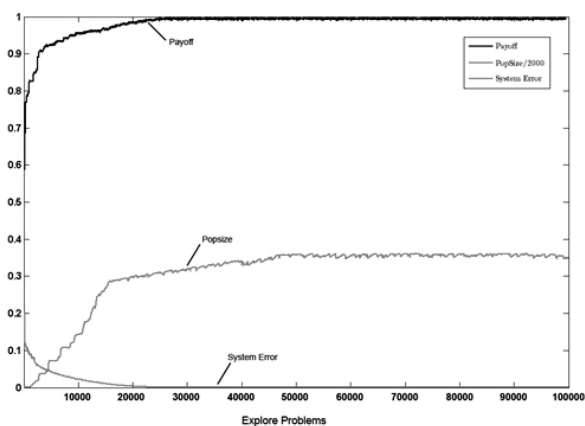


Figure 3: Results of *frog1* average over ten runs

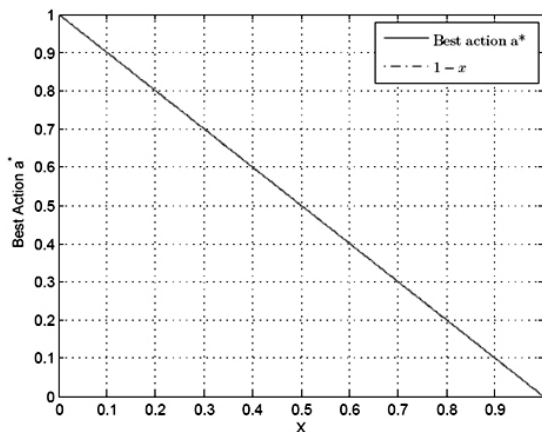


Figure 4: Best action  $a^*$  of *frog1* is plotted by scanning the values of  $x$  from 0 to 1, increased by 0.001 averaged over ten runs. Best action  $a^*$  is slightly broken at some input  $x$

Fig. 5 shows the best action  $a^*$  of *frog1* which is very similar to diagonal  $1-x$ . The *frog* problem seems simple, however the mapping from real distance  $d$  to transformed distance  $x$  makes it as a benchmark because this mapping from  $d$  to  $x$  hide the real position of fly and the frog-like system to be compelled to catch the fly using the payoff values. In the finalized population, the vectors of action computation coefficients  $v_0, v_1$  and  $v_2$  are close to 1, -1 and 0 respectively, and the standard deviations  $\sigma_0, \sigma_1$  and  $\sigma_2$  are very close to zero so in late phases the action computation coefficients  $v = (v_0, v_1, v_2)$  have little change.

Table1, shows 20 more activated classifiers from one run of *frog1*, where selected from the finalized population. In Table 1,  $l_0$  and  $l_1$  are the values for the interval predicate;  $w_0, w_1$  and  $w_2$  are the prediction weights for constant  $x_0$  and the input  $x$  and the action  $a$  respectively;  $v_0, v_1$  and  $v_2$  are the action computation coefficients for  $x_0, x$  and  $x^2$  respectively;  $\sigma_0, \sigma_1$  and  $\sigma_2$  are the standard deviations used by mutation on  $v_0, v_1$  and  $v_2$  respectively;  $\varepsilon, fit$  and  $num$  indicate prediction error, classifier fitness and numerosity respectively. For example in the last row of Table 1, classifier condition starts from 0.302 and spreads to 0.935, so from this interval  $x = 0.1$  and  $Cl.a$  is calculated by Eq.(13) as follows:  $Cl.a = v_2 * 0.01 + v_1 * 0.1 + v_0 = -0.0021 * 0.01 + -1.0023 * 0.1 + 0.9995 = 0.8992$

### 6.2. Frog2's results.

However *frog2* is more difficult than *frog1*; Fig. 6 shows the system performance, population size and system error of *frog2* where averaged over ten runs. The system performance is greater than 99% and the system error drops to smaller than 1% after 22000 explore problems. The final population size of classifiers  $N$  is smaller than 20% of  $N$  where it is about half of XCSFCA population size showed in [18]. Fig. 7 shows the best action  $a^*$  of *frog2* which is very similar to curve  $-\ln(x)$ .

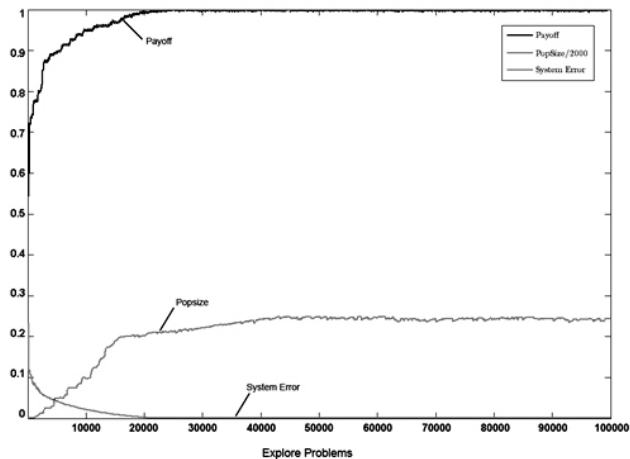


Figure 6: Results of *frog2* average over ten runs

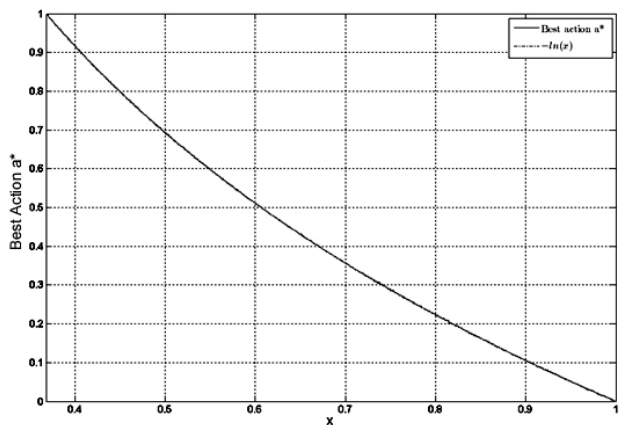


Figure 7: Best action  $a^*$  of *frog2* is plotted by scanning the values of  $x$  from 0 to 1, increased by 0.001 averaged over ten runs. Best action  $a^*$  is slightly broken at some input  $x$

Fig. 8 shows two classifiers with similar condition part where the first one calculate actions using linear function like XCSFCA and the second one calculate actions using polynomial equation. The condition part of both classifiers are  $Cl.c = [0.503, 0.812]$ , but the action weight vector of first classifier is  $Cl.\zeta = (\zeta_0, \zeta_1) = (-1.5210, 1.4251)$  and action computation coefficient of second one is  $Cl.v =$

$(v_0, v_1, v_2) = (1.90, -2.99, 1.11)$ . Fig. 8 clearly shows that the second classifier activity area is larger than first; consequently, in this type of action computation, the number of classifiers for handle a wide area of problem is smaller than linear case and the population tends to be compacted.

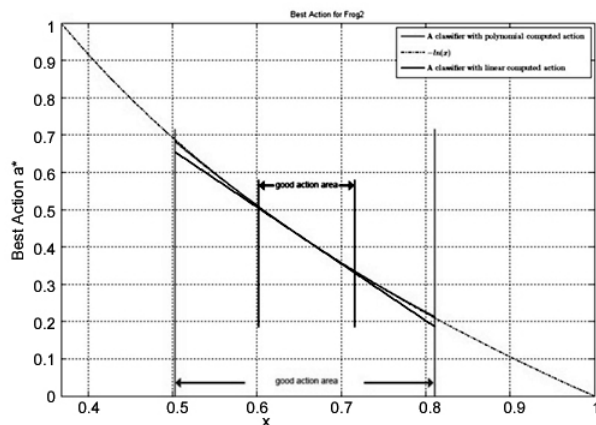


Figure 8: Compare two classifier activities; first classifier with polynomial computed action and second classifier with linear computed action

Table2, shows 35 more activated classifiers from one run of *frog2*, which are selected from the finalized population. All columns of Table 2 are same as Table 1. For example, in the first row of this table, the classifier condition starts from 0.398 and spreads to 0.706, so from this interval  $x = 0.4$  and  $Cl.a$  is calculated by Eq.(13) as follows:  $Cl.a = v_2 * 0.16 + v_1 * 0.4 + v_3 = 1.6209 * 0.16 + -3.6205 * 0.4 + 0.1.6209 = 0.9137$  while  $\ln(0.4) = 0.09163$ .

### 6.3. Frog3's results.

In Section 5.1, we described that both *frog1* and *frog2* are uniformly descending but *frog3* is not uniformly descending and has two direction changes. Calculating action by linear function is effective for *frog1* and *frog2* and similar problems because each classifier calculates the actions as a line, but for *frog3* linear function is not effective and need more classifier to cover all points of problem landscape.

Fig. 9 shows the system performance (in black), population size (in blue) and system error (in red) of *frog3* where averaged over ten runs. The system performance is greater than 99% and the system error drops to smaller than 1% after 29000 explore problems. The final population size of classifiers  $N$  is smaller than

30% of  $N$ . Fig. 10 shows the best action  $a^*$  of *frog3* which is very similar to curve  $\frac{1}{2}(\sin(2\pi x) + 1)$ .

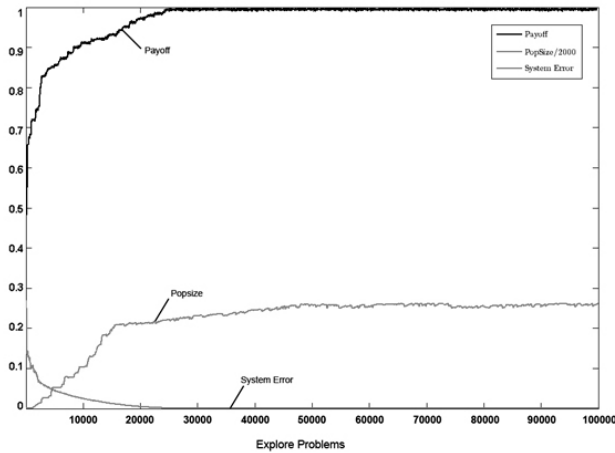


Figure 9: Results of *frog3* average over ten runs

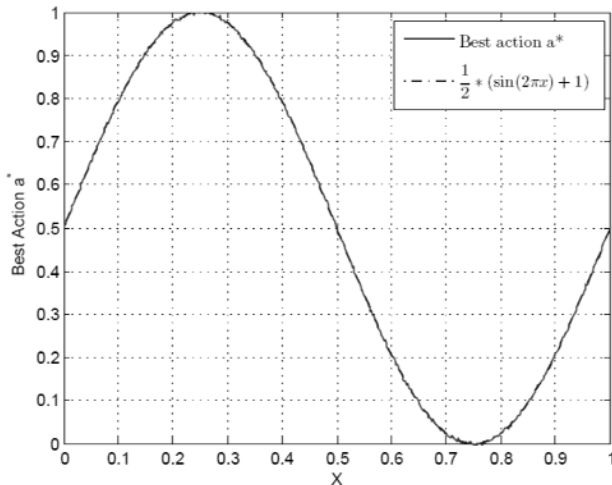


Figure 10: Best action  $a^*$  of *frog3* is plotted by scanning the values of  $x$  from 0 to 1, increased by 0.001 averaged over ten runs. Best action  $a^*$  is slightly broken at some input  $x$

Fig. 11 shows a classifier activity for *frog3* problem where selected from the finalized population of one run. In Fig. 11, the classifier condition is  $Cl.c = [l_0, u_1] = [0.105782, 0.54728]$ , the classifier actions are calculated by Eq. (13) where  $x$  starts from  $l_0$  to  $u_0$  and increases by 0.001 and the action computation coefficients is  $v = (v_0, v_1, v_2) = (0.47778, 4.089679, -8.149012)$ . Fig. 11 clearly shows that computing action using polynomial equation makes classifiers that are more powerful and in more complex problems such as *frog3*, XCSFCA could solve problem with smaller population size.

Table 3, shows 45 more activated classifiers from one run of *frog3*, that are selected from the finalized population. All columns of this table are same as Table 1. Fig. 12 shows the activity of classifiers showed in Table 3. Fig. 12

demonstrates that all points of curve  $\frac{1}{2}(\sin(2\pi x) + 1)$  can covered by the classifiers with continuous action using polynomial equation and we can conclude that cover this curve by classifiers that use linear function to compute actions, is more difficult and consequently needs larger population.

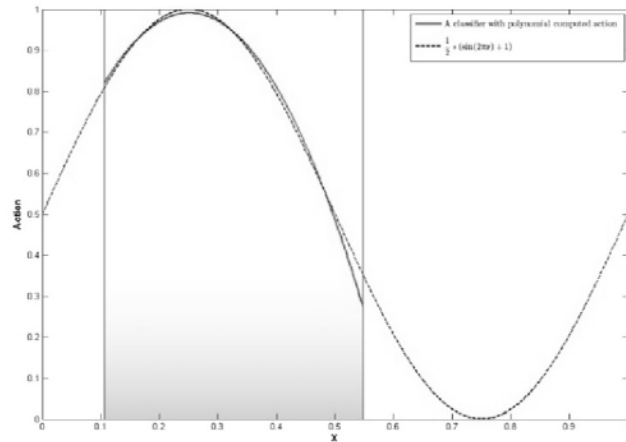


Figure 11: A classifier with polynomial computed action for *frog3* problem

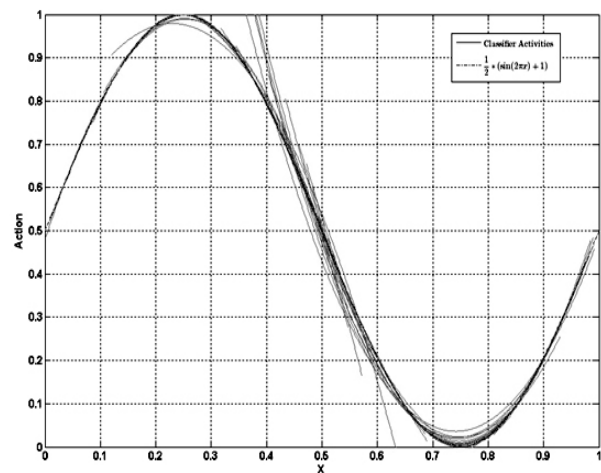


Figure 12: Classifier activities for *frog3* takes from table 3

## 7. Conclusion

In this paper we presented a new method for calculating continuous actions in which there would be directly computed as a polynomial equation of the input state  $a$  and a vector of polynomial coefficient  $Cl.v = (v_0, v_1, v_2)$ . We have shown that computing action using polynomial equation not only can solve the previous problems but it can solve more difficult problems with smaller population size. In our proposed method the classifiers is more

general because the mapping from input states to actions is none linear. Therefore, the action curve effectively can align to desired action curve. For better examination, we introduced a new *frog* problem called *frog3* where this is more difficult in compare to *forg1* and *frog2*.

Appendix

Table 1: Activated classifiers from one run of *frog1*

$l_0$	$l_1$	$w_0$	$w_1$	$w_2$	$v_0$	$v_1$	$v_2$	$\sigma_0$	$\sigma_1$	$\sigma_2$	$\epsilon$	fit	num
0.564	0.895	551.34	447.53	448.30	0.9435	-0.8359	-0.1133	0.0003	0.0002	0.0002	0.00003	0.932543	17
0.422	0.453	552.12	446.15	447.03	0.7595	0.1012	-1.2599	0.0001	0.0002	0.0001	0.00001	0.85019	34
0.512	0.696	551.15	446.17	446.39	0.8351	-0.4288	-0.4888	0.0002	0.0001	0.0002	0.00000	0.931921	26
0.405	0.724	550.76	449.89	449.99	1.0099	-1.0494	0.0562	0.0001	0.0000	0.0001	0.00002	0.936204	2
0.435	0.670	551.67	447.81	447.47	1.0783	-1.3053	0.2960	0.0001	0.0003	0.0002	0.00000	0.900447	39
0.309	0.817	551.34	446.71	445.91	0.9762	-0.8981	-0.1010	0.0003	0.0002	0.0001	0.00003	0.914114	20
0.351	0.834	550.17	448.96	448.56	0.9965	-0.9972	0.0016	0.0001	0.0003	0.0000	0.00003	0.912424	35
0.119	0.519	552.13	448.01	448.70	1.0094	-1.0726	0.1227	0.0001	0.0002	0.0002	0.00000	0.886267	32
0.053	0.506	550.81	448.93	448.68	1.0059	-1.0317	0.0339	0.0002	0.0000	0.0003	0.00001	0.933522	23
0.036	0.610	550.19	449.24	450.03	0.9957	-1.0011	0.0201	0.0003	0.0002	0.0003	0.00002	0.910299	18
0.014	0.751	550.37	449.38	448.45	0.9948	-0.9458	-0.0745	0.0001	0.0002	0.0002	0.00001	0.911272	14
0.406	0.847	550.46	449.51	448.80	1.0601	-1.2017	0.1624	0.0001	0.0002	0.0001	0.00000	0.886937	12
0.297	0.787	550.76	449.07	449.85	1.0262	-1.0952	0.0829	0.0000	0.0002	0.0003	0.00003	0.935824	7
0.591	0.965	550.24	448.99	449.85	0.9449	-0.8878	-0.0478	0.0002	0.0000	0.0001	0.00002	0.98569	36
0.125	0.782	550.18	448.87	449.22	1.0112	-1.0508	0.0465	0.0002	0.0002	0.0001	0.00000	0.859041	33
0.458	0.942	551.19	448.65	448.77	0.9300	-0.7961	-0.1428	0.0002	0.0000	0.0001	0.00002	0.869776	30
0.102	0.926	550.71	449.03	449.00	0.9942	-0.9912	-0.0026	0.0001	0.0000	0.0002	0.00001	0.923078	17
0.212	0.972	550.32	449.59	449.64	0.9925	-0.9505	-0.0465	0.0001	0.0000	0.0001	0.00001	0.918337	18
0.224	0.867	550.83	448.89	449.78	0.9887	-0.9548	-0.0390	0.0001	0.0003	0.0001	0.00001	0.863084	15
0.041	0.935	550.11	449.82	449.30	0.9995	-1.0023	-0.0021	0.0001	0.0001	0.0002	0.00000	0.934733	27

Table 2: Activated classifiers from one run of *frog2*

$l_0$	$l_1$	$w_0$	$w_1$	$w_2$	$v_0$	$v_1$	$v_2$	$\sigma_0$	$\sigma_1$	$\sigma_2$	$\epsilon$	fit	num
0.398	0.706	604.256	396.477	396.526	2.1026	-3.6205	1.6209	0.0003	0.0003	0.0000	0.0000	0.9537	9
0.391	0.629	585.297	414.412	413.953	2.1997	-4.0158	1.9946	0.0003	0.0000	0.0002	0.00002	0.8696	37
0.399	0.643	561.773	438.986	439.848	2.2198	-4.0811	2.0620	0.0001	0.0000	0.0001	0.00002	0.9758	28
0.405	0.500	590.845	409.476	410.112	2.3525	-4.6685	2.7014	0.0003	0.0000	0.0002	0.00002	0.8987	25
0.405	0.516	576.622	422.729	421.877	2.2265	-4.1553	2.1497	0.0001	0.0003	0.0003	0.00002	0.8611	9
0.420	0.536	612.416	387.540	388.513	2.3077	-4.4925	2.5071	0.0001	0.0001	0.0003	0.00003	0.9282	27
0.420	0.629	579.191	420.497	419.617	2.1635	-3.8855	1.8803	0.0002	0.0001	0.0001	0.00003	0.9713	50
0.446	0.742	639.625	360.915	360.770	1.9903	-3.2264	1.2854	0.0003	0.0002	0.0002	0.00001	0.9022	43
0.479	0.791	588.738	411.101	410.245	1.9896	-3.2382	1.2924	0.0001	0.0003	0.0001	0.00002	0.8521	2
0.480	0.799	600.556	399.426	398.663	1.9768	-3.1913	1.2548	0.0002	0.0001	0.0002	0.00000	0.9100	16
0.490	0.608	572.706	427.506	428.083	2.8821	-6.5533	4.3381	0.0003	0.0001	0.0001	0.00002	0.9063	28
0.503	0.812	598.578	401.457	401.738	1.9052	-2.9885	1.1108	0.0002	0.0001	0.0003	0.00001	0.9657	48
0.530	0.759	625.036	374.677	374.576	1.9234	-3.0317	1.1385	0.0001	0.0000	0.0002	0.00001	0.9430	49
0.530	0.783	598.628	402.074	402.496	2.0107	-3.3034	1.3426	0.0003	0.0000	0.0002	0.00003	0.9363	4
0.531	0.838	630.043	369.081	369.800	1.9341	-3.0842	1.1761	0.0001	0.0000	0.0001	0.00003	0.9007	9
0.541	0.849	596.694	403.953	404.266	1.9566	-3.1262	1.2020	0.0000	0.0003	0.0001	0.00002	0.9191	31
0.546	0.740	613.995	386.354	385.797	1.8822	-2.9157	1.0560	0.0001	0.0002	0.0002	0.00001	0.8973	34
0.548	0.663	628.909	370.247	371.084	1.8607	-2.8304	0.9648	0.0003	0.0000	0.0003	0.00002	0.9085	37
0.566	0.693	648.515	350.579	349.597	1.9129	-3.0237	1.1376	0.0003	0.0001	0.0001	0.00001	0.9135	13
0.568	0.993	602.391	397.975	397.244	1.8327	-2.7698	0.9414	0.0003	0.0001	0.0001	0.00001	0.9651	13
0.574	0.980	611.149	389.761	389.652	1.8068	-2.7000	0.9014	0.0001	0.0001	0.0002	0.00002	0.9418	69
0.579	0.939	627.030	373.264	373.685	1.8066	-2.7032	0.9022	0.0003	0.0001	0.0003	0.00001	0.9967	6
0.584	0.763	585.835	413.546	412.997	2.0794	-3.4752	1.4639	0.0000	0.0000	0.0002	0.00002	0.9611	45
0.602	0.748	655.315	345.240	346.014	1.9172	-3.0460	1.1553	0.0001	0.0002	0.0001	0.00001	0.9282	33
0.608	0.987	569.095	431.738	431.644	1.7512	-2.5447	0.8046	0.0003	0.0002	0.0002	0.00002	0.9981	5
0.613	0.757	573.704	427.036	427.983	1.8813	-2.9274	1.0793	0.0002	0.0002	0.0002	0.00002	0.9036	8
0.644	0.979	578.538	421.785	421.784	1.6683	-2.3524	0.6925	0.0000	0.0000	0.0002	0.00001	0.9886	10
0.657	0.863	621.729	377.725	378.440	1.8960	-2.9551	1.0741	0.0001	0.0001	0.0001	0.00001	0.9741	19
0.668	0.942	594.465	404.623	404.831	1.6236	-2.2387	0.6194	0.0002	0.0003	0.0003	0.00001	0.9227	26
0.670	0.845	615.929	383.544	384.027	1.6384	-2.2888	0.6479	0.0002	0.0002	0.0002	0.00000	0.9118	53
0.681	0.991	587.441	411.934	412.173	1.6209	-2.2347	0.6040	0.0000	0.0000	0.0001	0.00001	0.9442	21
0.683	0.871	644.612	354.853	354.850	1.6956	-2.4515	0.7647	0.0002	0.0001	0.0001	0.00001	0.8855	25
0.738	0.831	630.146	370.818	371.600	0.5328	0.5937	-1.2293	0.0002	0.0002	0.0002	0.00001	0.9312	17
0.838	0.992	601.617	398.968	399.775	-0.5003	2.4174	-1.9341	0.0002	0.0002	0.0002	0.00003	0.8864	19

Table 3: Activated classifiers from one run of *frog3*

$l_0$	$u_0$	$w_0$	$w_1$	$w_2$	$v_0$	$v_1$	$v_2$	$\sigma_0$	$\sigma_1$	$\sigma_2$	$\epsilon$	fit	num
-------	-------	-------	-------	-------	-------	-------	-------	------------	------------	------------	------------	-----	-----

0.236866	0.378151	557.448	442.355	442.105	0.462892	4.367334	-8.863423	0.0001	0.0000	0.0003	0.00001	0.9021	23
0.396261	0.504303	610.746	389.721	390.664	2.585839	-5.64389	2.983449	0.0001	0.0001	0.0001	0.00000	0.9256	6
0.675294	0.813556	565.184	435.705	435.391	5.199552	-13.8721	9.253198	0.0003	0.0002	0.0002	0.00003	0.8860	28
0.144715	0.270772	610.681	390.154	389.984	0.400498	4.772644	-9.499598	0.0001	0.0000	0.0000	0.00001	0.9055	6
0.753308	0.887874	616.630	384.009	384.940	4.652703	-12.5434	8.450933	0.0001	0.0001	0.0002	0.00002	0.9040	32
0.908448	0.957071	595.380	404.479	405.128	-4.055715	6.486184	-1.952416	0.0002	0.0001	0.0002	0.00001	0.9255	21
0.220638	0.529526	637.686	363.275	362.487	0.685409	2.952123	-6.676628	0.0001	0.0000	0.0001	0.00002	0.9854	32
0.590939	0.615211	597.901	402.838	402.149	-0.875622	6.102896	-7.165173	0.0000	0.0002	0.0001	0.00002	0.9498	6
0.476973	0.494184	546.631	452.399	451.762	1.321259	-0.04879	-3.190553	0.0001	0.0002	0.0001	0.00003	0.9644	13
0.429482	0.689566	519.450	480.683	481.642	2.907427	-6.48469	3.3197	0.0001	0.0002	0.0003	0.00001	0.9254	26
0.001792	0.285244	502.017	497.264	497.112	0.479286	3.928016	-7.350406	0.0001	0.0000	0.0002	0.00002	0.8668	16
0.434314	0.989304	614.745	384.941	384.918	4.518964	-12.053	8.060292	0.0001	0.0001	0.0002	0.00003	0.9397	18
0.810792	0.923059	511.993	488.662	488.867	2.549828	-7.67451	5.631641	0.0000	0.0001	0.0002	0.00000	0.9816	7
0.386302	0.975958	556.146	443.034	442.205	4.26313	-11.4037	7.658942	0.0002	0.0002	0.0002	0.00001	0.8891	4
0.105782	0.54728	599.462	401.290	401.297	0.47778	4.089679	-8.149012	0.0002	0.0000	0.0002	0.00003	0.9141	34
0.038944	0.572351	547.421	451.933	451.841	0.46726	4.11836	-8.122994	0.0002	0.0001	0.0003	0.00000	0.9989	12
0.242885	0.454341	625.044	374.427	375.312	0.548812	3.80329	-7.960753	0.0000	0.0001	0.0002	0.00001	0.9415	9
0.884585	0.896512	581.338	417.682	416.684	-2.485004	3.577624	-0.653736	0.0002	0.0002	0.0002	0.00001	0.9676	2
0.008574	0.107252	539.416	461.281	461.744	0.49494	3.389111	-3.951048	0.0001	0.0001	0.0002	0.00001	0.9781	16
0.59412	0.841014	620.532	379.419	378.934	5.044566	-13.4136	8.920132	0.0002	0.0001	0.0002	0.00003	0.9960	18
0.482966	0.912977	616.104	384.263	385.009	4.786066	-12.7864	8.550762	0.0001	0.0001	0.0000	0.00001	0.9445	33
0.21136	0.553078	588.525	411.761	412.337	0.717369	2.735222	-6.337924	0.0001	0.0003	0.0002	0.00002	0.9211	8
0.769896	0.847555	566.054	434.285	434.793	2.344688	-6.86859	4.962867	0.0002	0.0001	0.0002	0.00000	0.9155	3
0.41739	0.500036	577.489	422.148	422.633	4.848799	-15.4056	13.485216	0.0001	0.0000	0.0002	0.00003	0.9217	15
0.295148	0.352	597.111	403.392	404.261	0.387123	4.873423	-9.704807	0.0001	0.0001	0.0000	0.00003	0.8977	23
0.378401	0.93028	535.291	463.994	464.410	4.047738	-10.7379	7.162291	0.0003	0.0001	0.0001	0.00003	0.8694	29
0.316721	0.991723	552.987	446.608	446.387	3.739284	-9.98436	6.73005	0.0001	0.0003	0.0001	0.00001	0.9061	8
0.403481	0.80944	541.112	458.978	459.369	3.566391	-8.98479	5.635987	0.0002	0.0001	0.0001	0.00001	0.9364	31
0.356699	0.986483	542.287	456.866	456.244	4.085368	-10.9451	7.369765	0.0000	0.0003	0.0001	0.00002	0.9090	24
0.120854	0.638617	551.996	448.509	448.036	0.660247	2.776901	-6.040679	0.0002	0.0002	0.0000	0.00001	0.9108	35
0.382687	0.412435	584.311	415.898	415.093	1.397067	-0.47681	-2.57908	0.0000	0.0001	0.0002	0.00001	0.9675	1
0.630326	0.984551	556.056	443.997	444.666	4.810479	-12.8266	8.55884	0.0002	0.0001	0.0001	0.00001	0.9813	13
0.471696	0.9234	582.284	417.544	417.768	4.760066	-12.7141	8.502911	0.0002	0.0001	0.0001	0.00000	0.9128	31
0.140224	0.353918	577.498	423.425	424.000	0.405069	4.754791	-9.51792	0.0000	0.0003	0.0001	0.00002	0.8532	36
0.540241	0.645817	607.457	392.884	391.971	3.705055	-9.16138	5.55151	0.0002	0.0003	0.0002	0.00001	0.9645	24
0.248794	0.332006	611.469	388.959	389.148	0.471754	4.298072	-8.729586	0.0002	0.0002	0.0002	0.00003	0.9836	4
0.086551	0.407214	575.873	424.827	425.413	0.420461	4.625187	-9.249066	0.0001	0.0000	0.0002	0.00001	0.9079	19
0.816528	0.901481	619.306	380.119	379.261	4.270738	-11.6626	7.94208	0.0001	0.0003	0.0003	0.00002	0.9962	29
0.200032	0.314307	591.956	408.027	408.300	0.416822	4.701121	-9.464894	0.0003	0.0003	0.0003	0.00002	0.9158	29
0.345596	0.566242	527.571	472.094	472.440	1.420445	-0.52537	-2.617408	0.0001	0.0002	0.0001	0.00002	0.9697	35
0.347297	0.562507	590.386	408.891	409.880	1.405706	-0.4676	-2.673811	0.0002	0.0001	0.0000	0.00002	0.9226	5
0.31763	0.524651	498.064	502.112	501.684	1.098693	0.95305	-4.299129	0.0001	0.0001	0.0001	0.00002	0.9342	38
0.177281	0.213103	582.993	417.427	416.593	0.675133	1.941115	-2.197091	0.0003	0.0000	0.0000	0.00002	0.9527	33
0.478104	0.686496	568.397	432.095	432.336	3.423344	-8.2471	4.812014	0.0002	0.0001	0.0002	0.00001	0.8610	20
0.45901	0.945764	519.999	479.939	479.635	4.631968	-12.3687	8.273945	0.0000	0.0003	0.0002	0.00000	0.9122	18

## References

1. L.Bull, A.Sha'Aban,A. Tomlinson,J. Addison, and B.Heydecker,"Towards distributed adaptive control for road traffic junction signals using learning classifier systems", in Applications of LCS,Studies in fuzziness and soft computing,2004, pp. 276–299.
2. M.V.Butz,"An algorithmic description of ACS2", inLanzi PL, Stolzmann W, Wilson SW (eds) advances in learning classifier systems. LNAI,2002, Vol. 2321, pp. 211–229
3. M. V.Butz"Kernel-based, ellipsoidal conditions in the real valuedXCS classifier system", In Beyer HG, O'Reilly UM (eds) Genetic and evolutionary computation conference, GECCO, 2005. pp. 1835–1842.
4. M. V. Butz, Rule-based evolutionary online learning systems, Berlin: Springer,2006.
5. M. V. Butz,P. L.Lanzi, andS. W. Wilson,"Hyper-ellipsoidal conditions in XCS: rotation, linear approximation, and solution structure", inGECCO, 2006, Vol. 8, pp. 1457–1464.
6. M. V. Butz,andS. W. Wilson,"An algorithmic description of XCS", inAdvances in learning classifier systems, LNAI, 2001, Vol. 1996,pp. 253–272.
7. M. Dorigo,"Alecsys and the autonomous: learning to control a real robot by distributed classifier systems", in Mach Learn,1995, Vol. 19, pp. 209–240.
8. M. Dorigo, andM.Colombetti,"Robot shaping: an experiment in behavior engineering",Massachusetts: MIT Press/Bradford Books, 1998.
9. M.Dorigo, andU.Schnepf,"Genetics-based machine learning and behavior based robotics" in a new synthesis. IEEE Trans Syst Man Cybern, 1993, Vol. 23, pp. 141–154.
10. A. E. Eiben and J. E. Smith, Introduction to Evolutionary Computing, Springer, 2003.
11. P. L.Lanzi,"Learning classifier systems: then and now". Evol. Springer,2008,pp. 63–82.
12. P. L.Lanzi,D.Loiacono,S. W. Wilson,D. E. Goldberg, "Prediction update algorithms for XCSF: Rls, kalman filter, and gain adaptation", in GECCO, Vol. 8, 2006, pp. 1505–1512.
13. P. L.Lanzi,D.Loiacono,S. W. Wilson,D. E. Goldberg,"Generalization in the XCSF classifier system: analysis, improvement, and extension",EvolComput J,Vol. 15, No.2,2007, pp. 133–168.
14. D. Loiacono, andP. L.Lanzi,"XCSF with neural prediction",in IEEE congress on evolutionary computation. CEC 2006, pp. 2270–2276.
15. D.Loiacono,A.Marelli, andP. L.Lanzi, "Support vector regression for classifier prediction",in GECCO,2007, Vol. 2, pp. 1806–1813.
16. W.Stolzmann,and M. V.Butz, "Latent learning and action planning in robots with anticipatory classifier systems", inlearning classifier systems, from foundations to applications, Lecture notes in computer science,2000,Vol. 1813, pp. 301–320.
17. R. S.Sutton, andA.G.Barto, "Reinforcement learningan introduction",Cambridge: MIT Press, 1998.
18. T. H.Tran,C.Sanza,Y.Duthen, andT. D. Nguyen, "XCSF with computed continuous action", inGECCO, 2007, pp. 1861–1869.
19. B.Widrow, and M. E. Hoff."Adaptive Switching Circuits", Chapter Neurocomputing: Foundation of Research,Cambridge: The MIT Press, pp. 126-134, 1998.
20. S. W.Wilson, "Classifier fitness based on accuracy",EvolComputVol.3, No.2, pp.149–175,1995.
21. S. W.Wilson, "Get real! XCS with continuous-valued inputs", in Lanzi PL, Stolzmann W, Wilson SW (eds) Learning classifier systems, from foundations to applications, Lecture notes in computer science,2000, Vol. 1813, pp. 209–222.
22. S. W. Wilson,"Function approximation with a classifier system",in GECCO, 2001, pp. 974–981.
23. S. W. Wilson, "Classifiers that approximate functions". J Nat Compute, Vol. 1, No. 2,2002, pp. 211–234.
24. S. W. Wilson,"Classifier Systems for Continuous Payoff Environments",in GECCO, 2004, pp. 824-835 in Part II.
25. S. W. Wilson,"Three architectures for continuous action", in :Kovacs T, Llorà X, Takadama K, Lanzi PL, Stolzmann W, Wilson SW (eds) IWLCS, Lecture notes in computer science,2005, Vol. 4399, pp. 239–257.

**S. Goodarzian** was born in Shiraz, Iran in 1984. He received his B.Sc. degree in Computer engineering from Shiraz Islamic Azad University in 2008. He is received his M.Sc. degree in Artificial Intelligence at Shiraz University in 2011. His research interests include evolutionary computation and learning classifier systems.

**A. Hamzeh** received his Ph.D. in artificial intelligence from Iran University of Science and Technology in 2007. Since then, he has been working as assistant professor in CSE and IT Department of Shiraz University. His research interests include evolutionary computation, optimization and learning classifier systems.

**S. Hashemi** received the PhD degree in Computer Engineering from the Iran University of Science and Technology, in conjunction with Monash University, Australia, in 2008. He is currently a lecturer in the Electrical and Computer Engineering School, Shiraz University, Shiraz, Iran. His research interests include data stream mining, database intrusion detection, dimension reduction, and adversarial learning.



# High-Performance Low-Power Digital Linear Interpolation Filter

Magdy El-Moursy<sup>1</sup>, Member IEEE and Ahmed G. Radwan<sup>2</sup>

<sup>1</sup> Mentor Graphics Corporation, Cairo, Egypt  
Electronics Research Institute, Cairo, Egypt

<sup>2</sup> Institute for Electronics Engineering, University of Erlangen, Nuremberg, Germany

## Abstract

A new technique to implement digital interpolation filter is presented in this paper. The technique employs a sample calculation functional block which reduces the hardware required to realize the filter by orders of magnitude. The filter is realized with 80X160  $\mu\text{m}^2$  using 65 nm CMOS technology. Over Sampling Rate of up to 256 is achieved for 16-bit digital data sampled at 705,600 bps. The filter dissipates 77.68 mW when operating at frequency of 833.3 MHz.

**Keywords:** Digital interpolation, Oversampling, Digital Low Pass Filters, DAC, Delta-Sigma Modulators, FFT, FIR, DSP

## 1. Introduction

Interpolation filter is used in many digital signal processing (DSP) applications. Delta-Sigma based Digital-to-Analog converters ( $\Delta$ - $\Sigma$  DAC) employ an up-sample and interpolation filter to obtain high precision by speeding up the conversion processing over less number of bits [1]-[3]. Shown in Fig. 1(a) is up-sampling and interpolation filter which is performed by inserting zeros (known as zero-stuffing) to the in-between over-sampled data samples followed by a digital Low Pass Filter (LPF). This technique is widely used to oversample/interpolate digital signal [2]-[4]. Digital low pass filters (LPF) require large computational blocks. Discrete Fourier Transform (DFT) and Fast Fourier Transform (FFT) are used to reduce the computational cost of the filter. Yet, FFT requires tens, if not hundreds, of multipliers to be implemented [5]-[10]. In addition to the high cost in terms of hardware, stability is another factor which adds to the complexity of using DFT [11]-[13]. In this paper a new multiplier-free technique, as shown in Fig. 1(b), to implement digital interpolation filter is presented. The technique could be used to interpolate digital data represented in either fixed point or floating point number format. The presented technique uses Finite Impulse Response (FIR) and requires much less hardware to realize the interpolation filter.

Oversampled digital data could be interpolated using different interpolation techniques. Hold interpolation is the simplest of all. The up-samples hold the value of the slow samples in the region between input samples. Up-sampling in  $\Delta$ - $\Sigma$  modulators is based on reducing the difference between the samples which makes hold interpolation less attractive (for  $\Delta$ - $\Sigma$  modulators). Non-linear interpolation requires storing input samples over long period to determine the over-sampled sequence. However, linear interpolation requires less hardware since only two samples are sufficient to determine the output sequence. Linear interpolation is sufficient when high Over-Sampling Rate (OSR) is used. Since widely used in DSP, linear interpolation is adopted in the proposed interpolation filter.

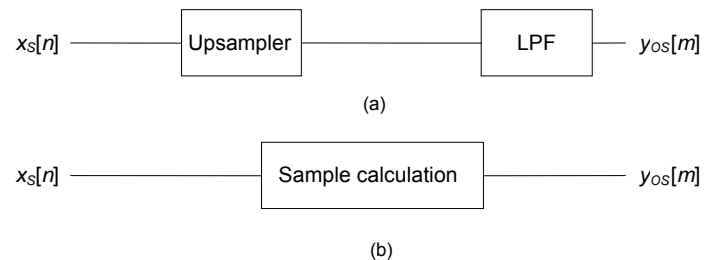


Fig. 1. Digital up-sample and interpolation filter. (a) conventional technique with LPF. (b) sample calculation technique.

The paper is organized as follows. In section 2, the theory behind the sample calculation of the interpolation filter is described. Some circuit implementation issues are described in section 3. In section 4, simulation results are presented. Some conclusions are provided in section 5.

## 2. Sample Calculation Techniques

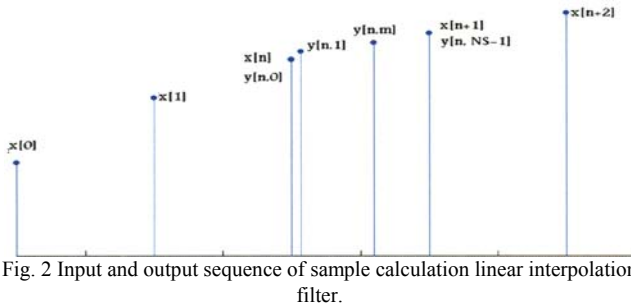
Sample calculation interpolation technique is presented in this section. In section 2.1, sample calculation interpolation technique for fixed point number representation is described.

The technique is extended to include floating point numbers in section 2.2. In section 2.3, the technique to evenly distribute the calculated samples is provided.

### 2.1 Fixed-Point Number Representation

Fixed point number representation is assumed to represent the input data samples  $x_s[n]$ , however, the methodology could be generalized to include both fixed point and floating point number representation as described in section II.B. Fixed point number representation dedicates  $k$  bits out of the total number of bits  $l$  to the integer part and  $(l-k)$  bits to the fraction part. Without loss of generality, the decimal point is shifted right  $(l-k)$  bits to convert the number to integer  $x[n]$  to simplify the sample calculation process. The decimal point of the output  $y[m]$  is shifted back left  $(l-k)$  bits to maintain the format of the number representation of the oversampled output  $y_{os}[m]$ .

Every two successive samples of the integer input data sequence  $x[n]$  and  $x[n+1]$  are used to determine the oversampled and interpolated output  $y[m]$ . The output  $y[m]$  (referred to by  $y[n,j]$ ) is produced at the  $OSR$ , where  $m=(n,j)$  is two dimensional index shown in Fig. 2,  $j$  is index changing from 0 to  $N_s-1$ , and  $N_s=OSR$ . In sample calculation interpolation filter, the value of the output sequence  $y[n,j]$  is determined by direct calculation of the samples to avoid using digital LPF which is needed in the conventional techniques. Since  $x[n]$  is integer,  $y[n,j]$  is expected to be an integer possessing the same precision (number of bits  $l$ ) of  $x[n]$ .



The difference between two successive integer (fixed point number with shifted decimal point) input samples  $x[n]$  and  $x[n+1]$ ,  $\Delta$  is integer as well,

$$\Delta = x[n+1] - x[n] \quad (1)$$

For ideal linear interpolation, the ideal output  $y_{id}[n,j]$  is determined by,

$$y_{id}[n,j] = x[n] + j \cdot Step \quad (2)$$

given

$$Step = \frac{\Delta}{N_s} \quad (3)$$

Since  $y[n,j]$  is integer and  $Step$  could be fraction, (2) could not be used to determine  $y[n,j]$ . The output  $y[n,j]$  could be either

$$y[n,j] = \lfloor y_{id}[n,j] \rfloor \text{ or } \lceil y_{id}[n,j] \rceil \quad (4)$$

Note that for perfect linear interpolation,

$$y[n,0] = x[n] \quad (5)$$

$$y[n,N_s-1] = x[n+1] \quad (6)$$

In order to satisfy (3)-(6), non-ideal interpolation is used. For DSP applications such as  $(\Delta-\Sigma)$  DAC, non-ideal interpolation is tolerable since the reduction in hardware complexity could be orders of magnitude. Furthermore, quantization error, always exists in DSP. Moreover, for high  $OSR$ , non-ideal interpolation is acceptable.

The number of output sequence samples  $y[n,j]$  in-between two successive input samples  $x[n]$  and  $x[n+1]$  equals  $N_s$ . In the ideal case each output sequence should be incremented by  $Step = \Delta/N_s$ . Rather than using fraction number to represent the output samples either the *ceiling* or the *floor* of the  $Step$  is used. In order to ensure that all output sequence samples are integer the output samples are divided into two sets of outputs  $y_c[n,j]$  and  $y_f[n,j]$ , where

$$y_f[n,j] = y[n,j-1] + N \quad (7)$$

$$y_c[n,j] = y[n,j-1] + N_1 \quad (8)$$

$$N = \lfloor \frac{\Delta}{N_s} \rfloor \quad (9)$$

$$N_1 = \lceil \frac{\Delta}{N_s} \rceil = N + 1 \quad (10)$$

The output sequence

$$y[n,j] = y_c[n,j] \text{ for } j = 0_{y_c} \quad (11)$$

$$= y_f[n,j] \text{ for } j = 0_{y_f}$$

The set of indices of the output samples which take the value  $y_c[n,j]$ ,  $0_{y_c}$ , and which take the value  $y_f[n,j]$ ,  $0_{y_f}$  is determined in section II.C.

In order to satisfy (5)-(8), the number of samples which take the value  $y_f[n,j]$ ,  $p$  and the number of samples which take the value  $y_c[n,j]$ ,  $q$  are to be determined to satisfy,

$$p + q = N_s \quad (12)$$

$$N \cdot p + N_1 \cdot q = \Delta \quad (13)$$

Given  $\Delta$  and  $N_s$ ,  $p$  and  $q$  are determined by solving (12) and (13) simultaneously,

$$p = N_1 \cdot N_s - \Delta \quad (14)$$

$$q = \Delta - N \cdot N_s \quad (15)$$

In order to reduce computational complexity in most DSP applications, *modulo-2 OSR* is used. Since only *floor* and *ceiling* of division by  $N_s$  are to be determined, simple  $r$ -bit left shift, where  $r = \log_2 N_s$ , is sufficient to realize (9) and (10). Also multiplication in (14) and (15) is realized by simple  $r$ -bit shift right, significantly reducing the hardware needed to calculate the samples and simplify the implementation of the samples calculation technique. Note that, no multipliers or divisors are needed to implement the filter.

### 2.2 Floating-Point Number Representation

In section II.A fixed point number representation was assumed for the input sequence  $x[n]$ . The presented technique

is generalized to include floating point numbers in this section. It is to be noted that equations (9), (10), (14), and (15) are the equations needed to realize the filter. The output samples are determined by simple addition in (7) and (8). Simple floating point addition/subtraction is needed to realize the filter. The output samples are determined by simple addition in (7) and (8). In addition, multiplication and division by  $N_s$  are to be implemented. Under the assumption of having *OSR modulo-2*, multiplication and division is performed on floating point numbers by simple addition and subtraction of the exponent of the number, respectively. Without lose of accuracy, in case of underflow or overflow in the exponent, the mantissa of the floating point number is treated like the integer number in section II.A. The technique could be used for both fixed point and floating point number representation without lose in accuracy. Once again multiplication and division are completely eliminated.

### 2.3 Uniform Distributed of Calculated Samples

It is shown in section II.A that the output sequence is divided into two sets of output samples; the samples which are incremented by  $N$ ,  $y[n,j]$  and the samples which are incremented by  $N_1$ ,  $y_c[n,j]$ . The number of output samples, which take the value  $y[n,j]$ ,  $p$  and the number of output samples, which take the value  $y_c[n,j]$ ,  $q$  are provided in (14) and (15) respectively. In order to obtain linear interpolation, the samples which take the values  $y[n,j]$  and  $y_c[n,j]$  are evenly distributed along the output samples sequence. Depending on solving (14) and (15), if  $p \neq q$ ; whether  $p > q$  or  $p < q$ , the majority of  $y[n,j]$  take either the value of  $y[n,j]$  or  $y_c[n,j]$ , respectively. Among the  $N_s$  output samples, majority samples  $y_{Maj}$  appears  $N_{Maj}$  times in the output samples and minority samples  $y_{Min}$  appears  $N_{Min}$  times in the output samples, where

$$N_{Maj} = \text{Max}(p, q), \quad (16)$$

$$N_{Min} = \text{Min}(p, q), \quad (17)$$

$$\text{if } N_{Maj} = p, \quad y_{Maj} = y[n,j], \quad (18)$$

$$\text{and } \quad y_{Min} = y_c[n,j], \quad (19)$$

$$\text{if } N_{Maj} = q, \quad y_{Maj} = y_c[n,j], \quad (20)$$

$$\text{and } \quad y_{Min} = y[n,j], \quad (21)$$

The distribution of the samples  $y_{Maj}$  and  $y_{Min}$  is based on one of two patterns *Patn1* or *Patn2* as shown in Fig. 3.

Sample index	0	1	2	...	...	...	<i>OSR-1</i>
Output sample	$y_{Maj}$	$y_{Min}$	$y_{Maj}$	...	...	...	$y_{Min}$

(a)

Sample index	0	1	2	...	...	...	<i>OSR-1</i>
Output sample	$y_{Maj}$	$y_{Maj}$	$y_{Maj}$	...	...	...	$y_{Maj}$

(b)

Fig. 3. The distribution of the minority and majority samples

in the output sequence a) pattern *Patn1* b) pattern *Patn2*.

Pattern *Patn1* is used when  $N_{Min} \geq N_{Maj}/2$ . Minority sample is assumed to exist after each majority sample in this pattern. The difference between the number of majority samples and minority samples,  $D = N_{Maj} - N_{Min}$  is determined. A number of minority samples equals  $D/2$  is replaced in the pattern *Patn1* with majority samples to satisfy (12), (16), and (17). The special case of  $p=q$  is included in this patterns when  $N_{Min}=N_{Maj}$ . In order to evenly distribute the majority and minority samples, the index of the minority samples which are replaced with majority samples,  $O_{y_{Maj}}(i)$  is determined by,

$$O_{y_{Maj}}(i) = 2 * \left\lfloor \frac{N_s}{2} (C + 1) \right\rfloor - 1 \quad (22)$$

where,  
 $C = D/2$

$$i = 1, 2, 3, 4, 5, \dots, C$$

For  $N_{Min} < N_{Maj}/2$ , pattern *Patn2* is used to distribute the samples. In *Patn2* all samples are assumed to be majority samples. A number of majority samples equals  $N_{Min}$  is replaced in the pattern with minority samples to satisfy (12), (16), and (17). In order to evenly distribute the majority and minority samples, the index of the majority samples which are replaced with minority samples,  $O_{y_{Min}}(i)$  is determined by,

$$O_{y_{Min}}(i) = 2 * \left\lfloor \frac{N_s}{2(N_{Min} + 1)} \right\rfloor - 1 \quad (23)$$

where ,  
 $i = 1, 2, 3, 4, 5, \dots, N_{Min}$

This distribution technique guarantees even distribution for the minority samples along the majority samples. The interpolated sequence is closest to linear under the digital nature of the output sequence. The linearity of the interpolation filter is demonstrated in section IV.

### 3. Circuit Implementation

The presented technique reduces the hardware required to realize the interpolation filter. Different techniques were considered to achieve further reduction in transistor count and power dissipation. A special case for the input sequence is to have no difference between two successive input samples ( $\Delta = 0$ ). In this case, all output samples  $y_{os}[m]$  take the value of the input sample  $x_s[n]$  reducing the power dissipation by eliminating the activity in the computation block. Also, division in (22) and (23) is implemented by iterative addition.  $N_s$  is divided by two first (before performing the division) to reduce the number of iterations of the division block.

In order to satisfy (12),  $D$  must be an even number.  $C$  is realized by single shift right operation. Binary shift and addition were also sufficient to realize (9), (14), and (15). Division used in calculating  $O(i)$  in (22) and (23) is done using addition and subtraction operation to calculate the floor value of the quotient. Some simulation results are included in

section IV.

## 4. Simulation Results

Some simulation results are provided in this section. In subsection IV.A, the interpolation algorithm is demonstrated. Simulation results for circuit implementation are summarized in subsection IV.B.

### 4.1 Interpolation Algorithm

The sample calculation interpolation algorithm is implemented using Matlab7.0. *OSR* of 64 is assumed. The over-sampled and interpolated output sequence is determined to demonstrate the accuracy of linear interpolation. Different input sequences are assumed. The output sequence is shown in Fig. 4 (a), (b), (c), and (d) for  $\Delta = 0$ ,  $N_{min} > N_{maj}/2$ ,  $N_{min} < N_{maj}/2$ , and  $p < q$ , respectively. Under the digital nature of the output sequence, highly linear and accurate interpolation is achieved as shown in the figure for different cases.

In Fig. 4(a),  $\Delta=0$ , consequently no interpolation is needed and the new samples have the same value of the original samples. In this case, both  $x[n]$  and  $x[n+1]=100$ ,  $p=64$ ,  $q=0$ , and consequently  $p$  is assigned to  $N_{Maj}$  and  $q$  to  $N_{Min}$ . In Fig 4(b),  $p > q$ ; the number of samples incremented by  $N$  is greater than that incremented by  $N_I$ . Also,  $N_{Min} \geq N_{Maj}/2$ ,  $Patn1$  is used to calculate the index of the output samples  $U_{y[n]}$ . For  $x[n]=50$  and  $x[n+1]=20$ ,  $p=34$ ,  $q=30$ ,  $p$  is assigned to  $N_{Maj}$  and  $q$  to  $N_{Min}$ . In Fig 4(c),  $p > q$ ,  $N_{Min} < N_{Maj}/2$ ,  $Patn2$  is used to calculate the index of the output samples  $U_{y[n]}$ . In this case  $x[n]=50$ ,  $x[n+1]=120$ ,  $p=58$ , and  $q=6$ . Consequently  $p$  is assigned to  $N_{Maj}$  and  $q$  to  $N_{Min}$ . In Fig. 4(d),  $p < q$ ; the number of samples incremented by  $N_I$  is greater than that incremented by  $N$ , in this case  $x[n]=20$ ,  $x[n+1]=70$ ,  $p=14$  and  $q=50$ . Consequently  $q$  is assigned to  $N_{Maj}$  and  $p$  to  $N_{Min}$ . High linearity is demonstrated in different cases in Fig. 4.

### 4.2 Interpolation Filter

The new interpolation filter is implemented using 65 nm CMOS technology. The sample rate of compact disc digital audio system of 44.1 KHz is used. For 16-bit data sample, input data sequence of 705,600 bps is assumed. The *OSR* is determined for the required performance. Linear interpolation with *OSR* up to 256 is achieved with the implemented technique. The filter occupies an area of 80X160  $\mu\text{m}^2$ . SPICE simulation is used to estimate the power dissipation of the circuit. The filter dissipates 77.68 mW when operating at 833.3 MHz. The filter specifications are summarized in Table 1.

## 5. Conclusions

In A new technique to implement digital interpolation filter is presented. The technique employs a sample calculation functional block, avoiding using multipliers in the filter and reducing the hardware to implement the filter. The filter is realized with 80X160  $\mu\text{m}^2$  using 65 nm CMOS technology. Over Sampling Rate of up to 256 is achieved for 16-bit digital data sampled at

705,600 bps. The filter dissipates 77.68 mW when operating at frequency of 833.3 MHz. The presented sample calculation technique reduces the hardware required to realize the filter by orders of magnitude while achieving higher Over Sampling Rate as compared to FIR technique.

## References

- [1] Bernard Sklar, *Digital Communication Fundamentals and Application; Second Edition*, Prentice Hall, 2004.
- [2] Tzu-Chiek Kue; Kwentus, A.; Willson, A.N., "A Programmable Interpolation filter for Digital Communications Applications", *IEEE International Symposium on Circuits and Systems*, Vol. 2, pp.97-100, May 1998.
- [3] Udo Zolzer, *Digital Audio Signal Processing*, John Wiley & Sons, 1997.
- [4] Richard G. Lyons, *Understanding Digital Signal Processing; Second Edition*, Pearson Education, 2004.
- [5] John G. Proakis and Dimitris G. Manolakis, *Digital Signal Processing, Principles, Algorithms, and Applications; Fourth Edition*, Prentice Hall, 2007.
- [6] C. J. Pan, "A Low Power Digital Filter for Decimation and Interpolation using Approximation processing", *IEEE International Solid-State Circuits Conference*, pp 102-103, 439, February 1997.
- [7] H.K. Kwan, "High-Order Tunable Passive Digital Filters", *IEEE International Symposium in Circuits and Systems*, Vol. 2, pp II.700-II.703, May 2002.
- [8] I.R. Khan, M. Okuda, and R. Ohba, "New Designs of Frequency Selective FIR Digital Filters", *IEEE International Symposium on Circuits and Systems*, Vol. 4, pp. IV.185-IV.188, May 2003.
- [9] M. Vollmer and H. Kopmann, "A Novel Approach to an IIR Digital Filter Bank with Approximately Linear Phase", *IEEE International Symposium on Circuits and Systems*, Vol. 2, pp.II.512-II.515, May 2002.
- [10] M. Bhattacharya and T. Saramaki, "Allpass Structures for Multiplierless Realization of Recursive Digital filter", *IEEE International Symposium on Circuits and Systems*, Vol. 4, pp. IV 237-IV.240, May 2003.
- [11] Steven r.Norsworthy, Richard Schreier, Gabor c. Temes, "Delta-Sigma Data Converters Theory, Design, Simulation." *IEEE Press*, 1997.
- [12] Robert S.Balog, "Topics In DSP: Interpolation & Delta Sigma Quantization", *Prentice Hall*, 1996.
- [13] Peter Kiss, Jesus Arias, Dandan Li, and Vito Bocuzzi, "Stable High-Order Delta-Sigma Digital-to-Analog Converters", *IEEE transactions on Circuits and Systems*, Vol. 51, no.1, Jan 2004.

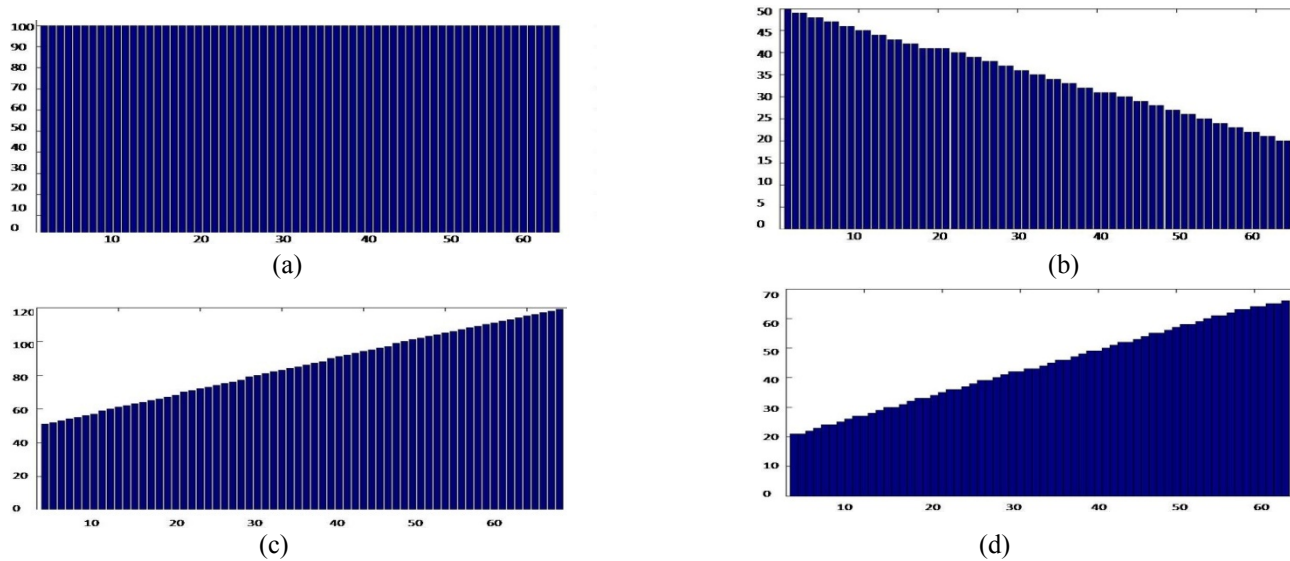


Fig. 4, Matlab simulation results for the interpolation filter. a)  $\Delta=0$ , no interpolation required, all output samples are equal . b)  $p>q$ ,  $N_{Min} > N_{Maj}/2$ ; *Patn1* is used. c)  $p>q$ ,  $N_{Min} < N_{Maj}/2$ ; *Patn2* is used. d)  $p<q$ .

**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 Staff Engineer 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.

**Ahmed Abdellatif** was born in Cairo, Egypt, in 1986. He received his B.Sc. from the German university in Cairo, Egypt in 2008, M.Sc. from the University of Ulm, Germany in 2010. From 2008 to 2009, he worked as a teaching assistant in the German university in Cairo. From 2009 to 2010, he worked as a student scientist in university of Ulm, on amplifiers for Retinal implants. Currently, he is working on his Ph.D. in Friedrich-Alexander University Erlangen-Nuremberg, where his main research point is designing MMIC's for broadband operation up to 100 GHz for radar and medical applications.

# Mobile Agent Based Hierarchical Intrusion Detection System in Wireless Sensor Networks

Surraya Khanum<sup>1</sup>, Muhammad Usman<sup>2</sup> and Ala'a Alwabel<sup>3</sup>

<sup>1</sup>Department of Computer Science & Information Systems, Faculty of Computer Science  
King Khalid University, Abha, Kingdom of Saudi Arabia

<sup>2</sup>Department of Computer Science, Faculty of Computer Science  
King Khalid University, Abha, Kingdom of Saudi Arabia

<sup>3</sup>Department of Information Systems, Faculty of Computer Science  
King Khalid University, Abha, Kingdom of Saudi Arabia

## Abstract

Security mechanism is a fundamental requirement of wireless networks in general and Wireless Sensor Networks (WSN) in particular. Therefore, it is necessary that this security concern must be articulated right from the beginning of the network design and deployment. WSN needs strong security mechanism as it is usually deployed in a critical, hostile and sensitive environment where human labour is usually not involved. However, due to inbuilt resource and computing restriction, security in WSN needs a special consideration. Traditional security techniques such as encryption, VPN, authentication and firewalls cannot be directly applied to WSN as it provides defence only against external threats. The existing literature shows that there seems an inverse relationship between strong security mechanism and efficient network resource utilization. In this research article, we have proposed a Mobile Agent Based Hierarchical Intrusion Detection System (MABHIDS) for WSN. The Proposed scheme performs two levels of intrusion detection by utilizing minimum possible network resources. Our proposed idea enhances network lifetime by reducing the work load on Cluster Head (CH) and it also provides an enhanced level of security in WSN.

**Keywords:** *Wireless Sensor Networks, Mobile Agent, Network Security, Intrusion Detection System, Hierarchical IDS.*

## 1. Introduction

Wireless Sensor Network (WSN) is an emerging technology [1,2]. The WSN is generally deployed in a critical and hostile environment where the human labour is not implicated. Some of the trendy applications of WSN are fire response, traffic monitoring, military command etc. [1,2,3,4].

Different types of network topologies such as star, tree, mesh etc are used for communication in WSN. In a cluster based hierarchical approach, concentration of sensor nodes forms a cluster and one node among them acts as a Cluster Head

(CH). The CH assumes to have a larger battery and acts as a supervisor node for communication between other nodes. All CH in the network are connected to a Base Station (BS) which is a single decision making authority. One of the cluster topologies is depicted in figure 1.

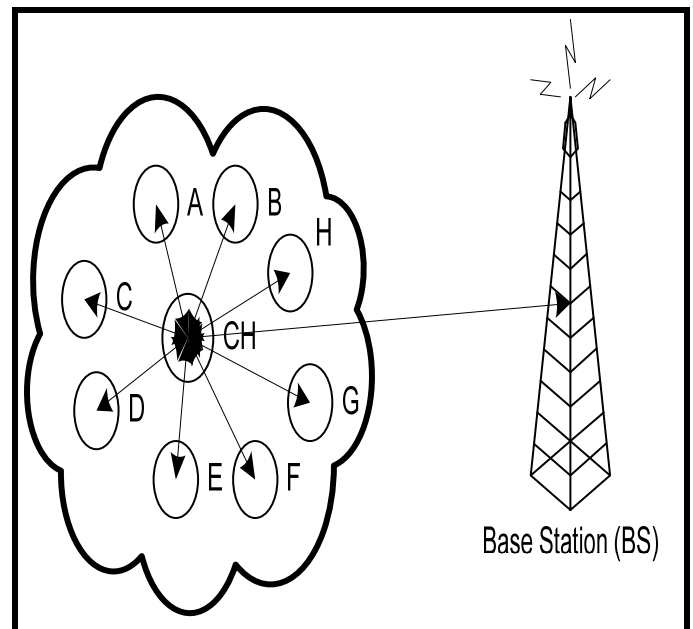


Fig. 1 Cluster in a sensor network.

The CH is a special sensor node with the specific tasks of receiving, processing, storing and forwarding data collected from the member nodes of that specific cluster. Each CH must be connected with Regional Head (RH) or BS, depending on the deployment scheme of the WSN. The RH

works very much like CH, but unlike CH, the RH connects different CH together. All the RH in the network is connected to BS which is a central governing and decision making authority. In a typical deployment of wireless sensor network, there are three tiers. At the top tier BS is deployed, RH and CH comes at middle tier, whereas the sensor node comes at lower most tiers.

Security is a major challenge in WSN networks. Deployment of adequate security mechanism in WSN is critical due to its resource restricted nature. Traditional security techniques such as encryption, VPN, authentication and firewalls are inadequate for WSN as they provide protection only against external threats and resource hungry in nature. Consequently more sophisticated techniques are required to monitor and discover intrusions in WSN as these networks are usually deployed in critical and real time application.

To protect the network against the intrusion customarily there are two types of defensive approaches. These are known as static and dynamic defensive approaches [3]. Firewalls, VPN, authentication and encryption are common examples of static technique. They only provide security from external threats and called as first line of defence. If a node inside the network is compromised, the whole security will be compromised. Therefore there is a need of better security mechanism that prevents the network from both internal & external threats.

An Intrusion Detection System (IDS) is a dynamic monitoring system used to identify, examine and observe violated activities. It discovers breach and illegal access to confidentiality, unavailability, authorization, authentication, integrity and network resources [4].

The related works shows that there exist a trade-off between better security mechanism and efficient resource utilization of sensor networks. If we increase network security we have to compromise on efficient resource consumption and vice versa. As a result, better security mechanisms is required that uses network resources efficiently. In order to tackle with this issue, we have proposed a Mobile Agent Based Hierarchical Intrusion Detection System (MABHIDS). Our proposed scheme uses minimum network resources by providing enhanced level of security.

The rest of the paper is organized as; in section II we have discussed the related work. Section III is about proposed scheme. We have discussed advantages of the proposed scheme in section IV. In section V, we have evaluated the performance of proposed scheme. In section VI we have concluded the contribution of this research paper. Finally, the list of references is given in section VII.

## 2. Related Work

We have divided the survey of existing literature into five categories. These categories are Domain Introduction, Architecture of WSN, Security issues on WSN, Different security mechanisms in WSN and Applying IDS to WSN.

In domain introduction we have surveyed different research papers regarding WSN introduction. The WSN is an autonomous network to monitor physical and environmental situation. A gateway incorporates WSN to the other wired/wireless networks. They are usually deployed for monitoring critical application such as structural monitoring for buildings and bridges, industrial machine monitoring, process monitoring, asset tracking etc.

Three types of network topologies are used in WSN: Star, Cluster tree and Mesh network. In star topology each sensor is directly connected to gateway. In cluster tree the sensor nodes form a tree structure and the higher node is connected to the gateway. The data flows from lower level of node to the higher level of node [3, 5]. In mesh network each sensor node is directly connected to other sensor node forming a net of interconnection link with each other.

The existing literature regarding WSN architecture shows that the sensor nodes have lack of common framework with no standardization in protocol for communication. There exist no interoperability mechanisms between two components of sensor nodes developed by different companies [5]. The sensor node is composed of battery so resource aware protocol architecture is needed for efficient communication. They use by default broadcast medium for communication which increases the risk of network congestion. Similarly, the authors have outlined the physical architecture, power management, commercially available sensor nodes and their characteristics in [6].

We have performed an in-depth survey regarding security issues in wireless sensor networks[7,8,9,10,11,12,13,14,15]. The authors argue that security is a major concern in any type of network particular in WSN. The WSN have resource restriction constraint i.e. limited energy, low computation capability, small memory, vulnerable to physical capture and insecure nature of wireless communication channel. All this limitation makes security in WSN a challenging issue.

The authors have outlined the basic security requirement, threat model and security attacks. They have divided the security issues into five categories: cryptography, key management protocols, security, routing, secure data aggregation and intrusion detection along with their advantages and shortcomings. The authors have discussed security requirement i.e. data authentication, data

confidentiality, data integrity, data freshness, self organization, availability, time synchronization, secure localization, scalability, availability, accessibility and flexibility. They have examined different types of security attacks i.e. Sybil, Denial of Service (DoS), physical, node replication, privacy violation and traffic analysis. The authors have provided basic guidelines and defensive measures against these types of attacks. They observed the DoS threats and layer wise security problems and argued that limited resources make encryption keys and digital signatures inadequate for securing WSN. Further, notify that there is a trade-off between energy and communication distance between sensor nodes therefore it should also be well managed.

Then we surveyed different existing security mechanisms in WSN. In [16] the authors assumed that base station is capable for storing all cryptography keys having sufficient memory and battery power. In [17] the author presents security architecture for mobile wireless sensor nodes by using a cryptographic algorithm. This algorithm proposes an authentication mechanism between the sensor nodes which provides security only from external threats.

The authors proposed a protocol called BROadcast Session Key Negotiation Protocol (BROSK) in [18]. This BROSK protocol uses broadcasting key negotiation message to provide link dependent keys to the sensor nodes for communication. This scheme uses simultaneous transmission for communication that increases the rate of collision.

In [19] the author proposed key distribution scheme using tree based approach in Wireless Sensor Network (WSN). They discussed scenario of sharing key when a new sensor node joins a network and assume to share its key with its neighbour. The proposed scheme is complex in nature and need extra computation resources. In [20] the authors have provided a framework for security with three management schemes for WSN. They evaluate these schemes on WSN challenging issues such as memory constraints, energy utilization, communication patterns, scalability, connectivity and communication patterns. The authors evaluate SACK, SACK-P and SACK-H management keys. The result shows that there exists inverse relationship between security and available resource utilization.

The above discussed security techniques such as encryption and authentication are not best suited in WSN environment. Therefore, there is a need for dynamic security mechanism in WSN. Intrusion Detection System (IDS) provides the dynamic security mechanism to WSN. We have surveyed several research articles in which variety of IDS are installed on WSN. Let's have a look at some of these techniques.

In [21] the authors have differentiated the available securities models. Currently two types of models are used for security: Intrusion Prevention (IP) and Intrusion detection (ID). IP uses authentication and firewalls for securing the boundaries of the network and ID uses some detection mechanism for identifying the intrusion in the networks. In [22] the authors notify the difference between IDS approaches for identifying and deflecting attacks. Host-based and Network-based are two types of approaches used by IDS. The authors highlighted the strengths and weakness of each approach. They argued that both of these techniques work together for achieving better intrusion detection and prevention.

Whereas in [23] the authors proposed a distributed intrusion detection scheme to monitor neighbour nodes for bringing the network back to function. They assume that adversary cannot capture or introduce new nodes inside the network. The proposed scheme creates a trust relation on neighbouring nodes which is not suitable if the trusted node is under attack.

Whereas, in [24] the authors have introduced a technique that observes the neighbourhood node communication called the spontaneous watchdog. The authors assume that the sensor nodes are stationary and used MICA2 radio stack for energy consumption. The decision for the selection of spontaneous watchdog imposed workload on the nodes and extra energy is required for activating global agent is the major drawback of this technique. In addition, the nodes are independent which do not assure only one global agent is activated per packet in the network.

The bottom-line is that, in existing IDS schemes in WSN, there seems an inverse relationship between enhanced security and efficient resource utilization in WSN. We need a better security mechanism which optimally utilize the resources of the WSN and provides better level of overall security.

### 3. Proposed Scheme

We have proposed a Mobile Agent Based Hierarchical Intrusion Detection System (MABHIDS) that provide two tiers of security in WSN. In this portion, we will discuss the architecture and working paradigm of proposed scheme in Section A and Section B respectively.

#### 3.1 Architecture

In order to provide two tiers of security we have installed Musk architecture [25] on each Cluster Head (CH). We have modified the MUSK architecture in order to behave as mobile agent. This architecture works as the Network Intrusion Detection System (NIDS) as well as Local



Intrusion Detection System (LIDS) on WSN. We have used two threshold frequencies. The threshold 1 is set on each CH for the normal activity of the network and threshold 2 is set on each sensor node for its normal activity.

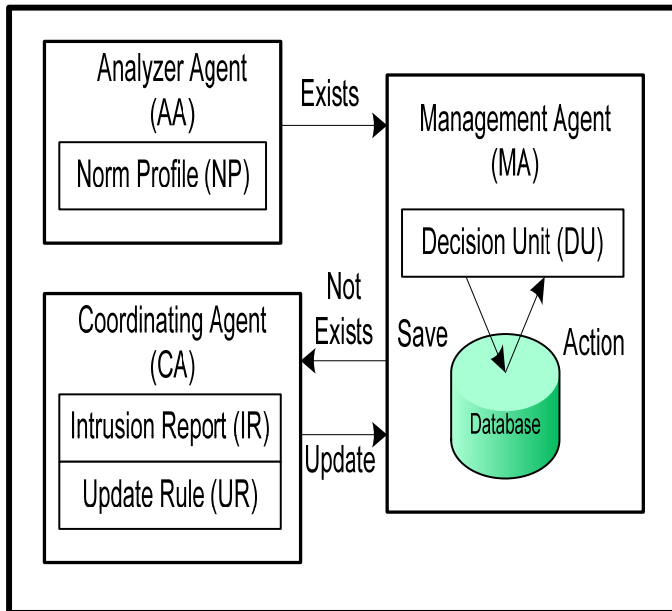


Fig. 2 MUSK Architecture [25]

**NIDS:** The different agents of Musk architecture [25] works as NIDS which is installed on each Cluster Head (CH) within a network. The NIDS capture the data packets along the path to identify an intrusion activity. The modified form of Musk architecture is shown in the Fig.2. This architecture is comprises of three agents: Analyzer Agent (AA), Coordinating Agent (CA) and Management Agent (MA). When the CH detects an intrusion it sends a copy of Analyzer agent (AA) to the victim node. Therefore AA is mobile in nature. The CA and MA are preset in the CH and they are fixed.

**Analyzer Agent (AA):** The Analyzer Agent (AA) is used to monitor node activity. It is a mobile agent and installed on each CH in the network. When CH discovers an intrusion it sends a copy of AA to the suspicious node. The AA uses victim resources in order to verify the occurrences of intrusion. The AA generates a Norm Profile (NP) and check the threshold 2. If there is a deviation from the threshold frequency the AA generates an alarm and notifies the CH. The CH calls the Management Agent (MA) for analysis.

**Management Agent (MA):** The Management Agent (MA) contains a sub unit called Decision Unit (DU) for the analysis of intrusion. The DU maintains the database of already occurred intrusions. When an intrusion occurs the

CH calls the MA for analysis. The MA activates its DU that searches in its database whether this intrusion happens in the past or not. The database contains the predefined stored intrusions along with the decisions. If the match occurs against the pre stored intrusions then DU performs already stored decision and informs to the CH. If there is no such entry in the database then MA informs the Co-ordinating Agent (CA) regarding the occurrence of novel intrusion.

**Coordinating Agent (CA):** The Coordinating Agent (CA) performs two basic functions i.e. generate Intrusion Report (IR) and Update Rule (UR). When CA receives a novel intrusion message from MA it sends to IR. The IR forwards this report to the Base Station (BS) regarding the occurrence of intrusion. The BS is a centralized decision making authority against the intrusion. It makes a decision on novel intrusion and sends it to the Update Unit (UU). The UU generates new rule against that intrusion and send it to MA. The MA saves the intrusion in the database for future use. If the same intrusion happens again the DU searches the database and performs the already stored decision.

**LIDS:** The Analyzer Agent (AA) is a mobile agent and works as LIDS. When NIDS in CH deviate from its threshold 1 it generate an alarm informing the occurrence of intrusion. The CH makes analysis and identifies the sensor node that is generating abnormal traffic. The CH activates its mobile AA and send to the victim node. The AA works as LIDS and uses resources of the suspicious node for identifying the malicious activities. The AA informs the CH either the suspicious node is victim or safe. If the node is victim the CH that takes appropriate action upon that activity. The copy of AA is only send to the suspicious node instead of installing LIDS on each sensor node.

The fig-3 is representing a working deployment of NIDS & LIDS. It is vital to mention here that the NIDS is deployed on each CH whereas the actual deployment of LIDS is also at CH. On each intrusion alarm, the LIDS (which are a mobile agent) are triggered by CH for further inspection of the behaviour of suspicious node. The LIDS uses resources of suspicious node to report it either as a victim or safe node.

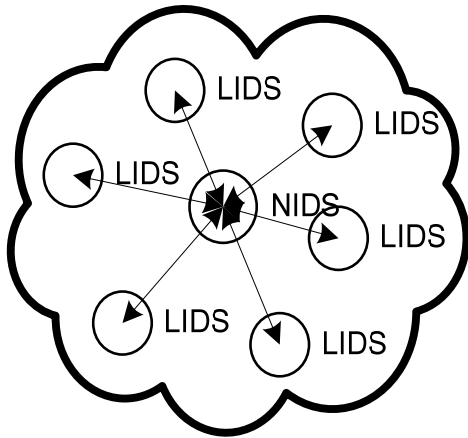


Fig. 3 Working Deployment of NIDS & LIDS.

### 3.2 Working Paradigm

We set two threshold levels for intrusion detection, one for Network Intrusion Detection System (NIDS) and other for Local Intrusion Detection System (LIDS). Threshold 1 is set on each CH over the network and works as the NIDS whereas; the threshold 2 is set on each sensor node and works as LIDS. The initial intrusion detection is performed by NIDS which detects the normal rate of packet arrival and departure. In case of deviation from threshold 1, the CH triggers the mobile Analyzer Agent (AA) over the link where deviation is occurred.

The AA will visit the suspicious node and acts as Local Intrusion Detector (LID) over there. The AA will use the resources of suspicious node to investigate its behaviour further. This investigation is based on threshold 2. If suspicious node is found the victim the AA will update CH. The CH inform its sub agents i.e. Coordinating Agent (CA) and Management Agent (MA) that will take appropriate action to prevent rest of the network from intrusion either by minimizing the communication with the victim node, reducing the trust value on the victim node or by cutting its communication from rest of the network. Otherwise, the AA informs the CH that the suspicious node is not the victim; it is a safe node and unusual but harmless activity has taken place.

### 4. Advantages

The Major advantage our proposed approach is that it provides two levels of security by using resources of sensor network optimally. It also reduces the workload of Cluster Head (CH) and provides enhanced security. As in existing schemes CH is responsible for all computation pertaining to the intrusion detection activity in member nodes of the CH. Whereas in our proposed scheme CH triggers the AA for suspicious node on its every unusual activity. The AA uses

suspicious node's resources in order to declare it either as a victim or safe node. In this way CH resources are saved as compare to the existing schemes. Another benefit of our approach is infrastructural reduction as we do not need to install LIDS on every node rather mobile agent acts as a LIDS on suspicious node. This enhances the overall life time of the sensor network.

### 5. Performance Evaluation

We have performed the analytical performance comparison of our proposed scheme with existing schemes. We analyzed their performance on two major factors i.e. Security and Efficiency.

The security factor is divided further into three parameters i.e. internal external and novel threats. Internal threats are those attacks that are initiated or injected by the intruder residing inside the network. External threats are from outside attackers. Novel threats are the unusual or unrecognized form of the intrusions which have not occurred previously. Three types of possible values used by these intrusions are low, high and medium that indicates how clearly the proposed scheme identifies these intrusions. We have given the low value to all those schemes that doesn't provide defence against the compromised node, under attack nodes, inside attackers, master or secret key is captured or the node activity is dependent on the neighbourhood node information, trust relationship on nodes etc. the medium value to the all those proposed scheme that identify the intrusion but does not provide any defensive measurement how to handle them, generate false negative in large amount. The high value to all those schemes that clearly identify the intrusion as well as provide the counter measure against that intrusion, compromise of one node will not make the whole security of the system vulnerable.

We divide the efficiency factor into three parameters i.e. computation costs, network bandwidth, node resource utilization and number of messages. Two types of values are used high and medium in computation cost, network bandwidth and node resource utilization. We have given high value to all those schemes that increases burden on network resource i.e. cryptographic algorithms are resource hungry in nature that require extra computation and memory overhead, communication steps between nodes increases, simultaneous transmission increases the rate of collision that effect the bandwidth issues, large amount of false negative dissipate the energy resources etc. The medium value is given to the scheme that uses victim resources in order to discover an intrusion by using minimum network resources. The number of messages which contains the integer value i.e. additional steps used by the proposed schemes in order to identify the intrusion. Table 1 shows that our proposed

scheme is efficient in several aspects as compare to the existing schemes

Table 1: Performance comparison between different existing schemes

Sr. No	Scheme Name	Security			Efficiency			
		Internal Threats	External Threats	Novel Threats	Comp. Costs	Network Band-Width	Node Resource	No. of Messages
1	Security Protocol for Sensor Networks [16]	Low	High	Low	High	High	High	8
2	A Security Architecture for Mobile WSN [17]	Low	High	Low	High	High	High	---
3	Scalable Session Key Construction Protocol for WSN [18]	Low	Low	Low	High	Medium	High	---
4	A Tree Based Approach for Secure Key Distribution in WSN [19]	---	---	---	High	Medium	High	4+4
5	A unified security framework with three key management schemes for WSN [20]	Low	High	Low	High	High	High	---
6	A Decentralized IDS for Increasing Security of WSN [26]	High	High	High	High	High	High	---
7	An IDS for WSN [27]	Medium	Medium	High	---	High	High	---
8	Anomaly Intrusion Detection in WSN [28]	Medium	Medium	---	---	---	---	---
9	Decentralized Intrusion Detection in WSN [29]	---	---	---	High	High	High	---
10	Intrusion Detection based Security Architecture for WSN [30]	Medium	Medium	High	High		High	---
11	Energy Efficiency of IDS in WSN [31]	High	High	High	High	High	High	---
12	An Improved IDS Based On Agent [32]	---	---	---	High	High	High	---
13	A Framework of Machine Learning Based Intrusion Detection for WSN [33]	---	---	Low	---	---	High	---
14	Mobile Agent Based Hierarchical IDS (Proposed Scheme)	High	High	High	Medium	Medium	Medium	---

## 6. Conclusions

The resource restricted nature of WSN demands a more sophisticated and secure security mechanism for these sorts of networks. There seems an inverse relationship in better security and optimum resource utilization of network resources in existing security schemes of WSN. In this research article, we have proposed a security model which not only provides good level of security but it also uses network resources optimally for the provision of better security. In proposed approach, we have proposed a two tier security model for WSN. The NIDS and LIDS are involved in providing two tier securities. The NIDS is installed on all CH whereas LIDS is based on mobile agent. The LIDS is activated whenever CH found any node suspicious. The CH issues LIDS for further scrutiny of malicious activities of suspicious node in order to affirm it as a compromised node. The LIDS uses resources of suspicious node. The proposed mechanism provides

enhanced security using resources of WSN optimally. The workload of CH is also reduced using our proposed. Our proposed approach also helps in security infrastructural reduction for enhanced security.

## References

- [1] S. Kaplantzis, "Security Models for Wireless Sensor Networks", Research Thesis, 2006.
- [2] A. Perrig, J. Stankovic and D. Wagner, "security in wireless sensor networks" communications of the ACM, vol. 47, no. 6, June 2004.
- [3] D. Culler, "Overview of sensor Networks" University of California, Berkeley Deborah Estrin Mani Srivastava University of California, Los Angeles, IEEE Computer society, August 2004.
- [4] A. Bob, "What is sensor network" National Instruments, LabVIEW, NI, White Paper.
- [5] F.L. Lewis, "Wireless Sensor Networks" Smart Environments: Technologies, Protocols, and Applications Conference, New York, 2004.

- [6] S. Ramesh, "A Protocol Architecture for Wireless Sensor Networks" School of Computing, University of Utah, 2006.
- [7] Y. Wang, G. Attebury, and B. Ramamurthy, "A survey of security issues in wireless sensor networks" University of nebraska-lincoln, IEEE Communications Surveys & Tutorials, 2006.
- [8] M. Sharifnejad, M. Sharifi, M. Ghiasabadi and S. Beheshti, "A survey on wireless sensor networks security" published in 4th international conference: sciences of electronic, technologies of information and telecommunications, Tunisia, 2007.
- [9] A.D. Wood and J.A. Stankovic, "Denial of Service in Sensor Networks", University of Virginia, IEEE Conference, 2002.
- [10] C-Y. Chong and S. P. Kumar, "Sensor Networks: Evolution, Opportunities, and Challenges" proceedings of the IEEE, vol. 91, no. 8, August, 2003.
- [11] A. Perrig, J. Stankovic, and D. Wagner, "Security In Wireless Sensor Networks", communications of the ACM, vol. 47, no. 6, June 2004.
- [12] Y. Wei, L. Paul and J.M. Havinga, "How to Secure a Wireless Sensor Network", Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, Netherlands, Published by IEEE ISSNIP, 2005.
- [13] T. Zia and A. Zomaya, "Security Issues in Wireless Sensor Networks" School of Information Technologies, University of Sydney, Published by IEEE, 2007.
- [14] R. Omer, O. Kasten and F. Mattern, "Middleware Challenges for Wireless Sensor Networks", Department of Computer Science, ETH Zurich, Switzerland, Mobile Computing and Communications Review, Volume 6, Number 2, 2004
- [15] R. Roman<sup>1</sup>, J. Zhou, and J. Lopez, "On the Security of Wireless Sensor Networks", Institute for Infocomm Research, Heng Mui Keng Terrace, Singapore and Ingenieria Informatica, University of Malaga, Malaga, Spain, 2005.
- [16] A. Perrig, R. Szewczyk, J.D. Tygar, Victorwen and E. Culler, "SPINS: Security Protocols for Sensor Networks" Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Wireless Networks, 2002.
- [17] S. Schmidt, H. Krahn, S. Fischer, and D. Watjen "A Security Architecture for Mobile Wireless Sensor Networks" by Springer-Verlag Berlin Heidelberg, LNCS, 2005.
- [18] B. Lai, S. Kim and I. Verbauwhede, "Scalable Session Key Construction Protocol for Wireless Sensor Networks", Department of Electrical Engineering University of California, Los Angeles, USA, 2003.
- [19] E. Blaß, M. Conrad and M. Zitterbart, "A TreeBased Approach for Secure Key Distribution in Wireless Sensor Networks.", 2006.
- [20] R. Riaz, A. Naureen, A. Akram, A.H. Akbar, K.H. Kim, H. F. Ahmed "A Unified Security Framework With Three Key Management Schemes For Wireless Sensor Networks", Elsevier, 17 June 2008.
- [21] J. G. Tront and R. C. Marchany, "Internet Security: Intrusion Detection & Prevention", IEEE Proceedings of the 37th Hawaii International Conference on System Sciences, 2004.
- [22] A. Bob, "Network- vs. Host-based Intrusion Detection" A Guide to Intrusion Detection Technology, ISS Internet Security System October 2, 1998
- [23] K. Ioannis, T. Dimitriou and F. Freiling, "Towards Intrusion Detection in Wireless Sensor Networks" Athens Information Technology, 19002 Peania, Athens, Greece and Department of Computer Science, University of Mannheim, Germany, 2006
- [24] R. Roman, J. Zhou and J. Lopez, "Applying Intrusion Detection Systems to Wireless Sensor Networks", Communications Society publication in the IEEE CCNC 2006.
- [25] S. Khanum, M. Usman, K. Hussain, R. Zafar, and Dr M. Sher, "Energy-Efficient Intrusion Detection System for Wireless Sensor Network Based on MUSK Architecture" HPCA 2009, LNCS 5938, pp. 212–217, Springer-Verlag Berlin Heidelberg 2010
- [26] I. Chatzigiannakis and A. Strikos, "A Decentralized Intrusion Detection System for Increasing Security of Wireless Sensor Networks", 2005.
- [27] I. Onat and A. Miri, "An Intrusion Detection System for Wireless Sensor Networks", published by IEEE, 2000.
- [28] V. Bhuse and A. Gupta, "Anomaly Intrusion Detection in Wireless Sensor Networks", Western Michigan University, Kalamazoo, USA, 2005.
- [29] Ana Paula et. al., "Decentralized Intrusion Detection in Wireless Sensor Networks", ACM, 2005.
- [30] D. Xiao, C. Chen and Gaolin Chen, "Intrusion Detection based Security Architecture for Wireless Sensor Networks", IEEE, 2005.
- [31] P. Techateerawat and A. Jennings, "Energy Efficiency of Intrusion Detection Systems in Wireless Sensor Networks", IEEE, 2006.
- [32] B. Dong and X-L. Liu, "An Improved Intrusion Detection System Based On Agent", IEEE, 2007.
- [33] Z. Yu and J.P. Tsai, "A Framework of Machine Learning Based Intrusion Detection for Wireless Sensor Networks" Department of Computer Science, University of Illinois at Chicago, IEEE, 2008.

## Authors:

**Surraya Khanum** is M.Phil / MS in computer science with the specialization of Computer Networks. She has numerous international publications in the area of network security. Currently she is working as a Lecturer in Computer Science in Faculty of Computer Science, King Khalid University, Abha, Kingdom of Saudi Arabia. Her research interests include networks, network security, wireless networks, sensor networks and intrusion detection systems.



**Muhammad Usman** is M.Phil / MS in Computer Science with specialization in Computer Networks. He is scholarship holder to undertake his PhD studies from Griffith University, Australia from Feb, 2012 to Jan 2015. He has 8 years of teaching, research and professional experience. Currently he is working as a Lecturer in Computer Science in King Khalid University, Abha, Kingdom of Saudi Arabia. He has several international publications. His research interest includes networks, network security, wireless networks, mobile ADHOC networks, wireless sensor networks, intrusion detection systems and ongoing issues in multimedia.

**Ala'a A. Alwabel** is MS in Information System with the specialization of information security. She has International publication in the area of Information Security. Currently she is working as a coordinator and lecturer in Faculty of Computer Science, department of Information System, King Khalid University, Abha, Kingdom of Saudi Arabia. Her research interests include Wireless Networks, Networks Security, Cryptography, Information Security, Information Retrieval and E-Health.

# New method to parse invoice as a type the document

MOUJABBIR Mohammed<sup>1</sup>, RAMDANI Mohamed<sup>2</sup>

<sup>1</sup> University Hasssan II Faculty of sciences and technology  
Department of computer system  
Mohammedia, BP 146 Mohammedia, Morroco

<sup>2</sup> University Hasssan II Faculty of sciences and technology  
Department of computer system  
Mohammedia, BP 146 Mohammedia, Morroco

## Abstract

In this paper We propose a new method able to detecting and correcting errors relating to the recognition of invoice type documents. We rely on automated document readers that can read and recognize the various relevant information in a scanned document.

The process on which this method is based consists of digitizing a large volume of documents, and makes them pass through automatic readers of the documents, then carry out the correction of the various errors. The final goal is to find an electronic document reflecting the various information included in the background document. The main goal is the generation of organized electronic documents, like a data basis or files XML; for a specific use.

Our approach is based on the language theory through developing a kind of parser which is applicable to the more general case of documents and can easily detect a specific class of errors and correct them.

**Keywords:** Documents dematerialization, electronic invoicing dematerialization, autorun, character automatic recognition, languages theory, compilation techniques.

## 1. Introduction

The current large development and deployment of dematerialization the documents has a great effect on the research activities in the domain of recognition documents, detection errors and also correction them.

Information and communication technologies have had considerable effects on how companies do business with their business partners [3]. In a narrow sense, these effects are based on electronic commerce (e-commerce), which is the buying of products from suppliers and their selling to customers using the new information technologies.

There are several models of e-commerce, namely, business-to-business, e-commerce between companies, business-to-consumer, e-commerce between companies and consumers and business-to-government, e-commerce between companies and government organizations.

The process of dematerialization is aimed at the transition from a physical document to an electronic document (structured document[1] or not) without human intervention [2]. This is made possible by using an OCR<sup>1</sup> and an error processing method. This method is crucial for such a transition which can't be, in any case, conducted in a transparent manner, i.e., without errors. Several injections of errors are due to several factors including printing quality, quality of the paper used, scanner resolution, software power (OCR) ... Hence, an error processing operation is necessary [1], that can be divided into two parts, one for detecting errors, and another to correct them.

The goal of this study is to provide a system assessing the different aspects especially relevant information contained in a physical document, into electronic textual contents.

The problem to be raised is to envisage a pretreatment of the errors [1], which ties to correct errors of recognition before reaching the learning phase. The existing solution will be presented: the arborescent[3] treatment. This solutions advance many anomalies which will be detailed in the following sections. However this paper introduces a new approach which will be compared to the XML technology, also they (approach) provide us the possibility to generate a new parser able to detect and correct errors. Practically this parser gives us very good results, specially when XML technology was not able to detect all the existing errors.

In this article we will detail the process of the document processing, with a focus on the arborescent method. In section III will present the anomalies and limitations generated by the existing method. The section VI reports the solution suggested which proposes a tally formal modeling; a language dedicated to the documents with a grammar and syntax. However the section V introduces the results obtained and a discussion of these

---

<sup>1</sup> : Optical character recognition.

results. The last section will focus on the practical implications of our study.

The text must be in English. Authors whose English language is not their own are certainly requested to have their manuscripts checked (or co-authored) by an English native speaker, for linguistic correctness before submission and in its final version, if changes had been made to the initial version. The submitted typeset scripts of each contribution must be in their final form and of good appearance because they will be printed directly. The document you are reading is written in the format that should be used in your paper.

This document is set in 10-point Times New Roman. If absolutely necessary, we suggest the use of condensed line spacing rather than smaller point sizes. Some technical formatting software print mathematical formulas in italic type, with subscripts and superscripts in a slightly smaller font size. This is acceptable.

## 2. The xml technology

XML(Extensible Markup Language) is becoming a dominant standard for storing and exchanging information. They use several tools such as DTD, XSLT,XSL-FO, XSLQuery, Schema-XML, each one of them has its own specification, and can be used in various domains like data warehousing ,web, e\_commerce .... Since XML is used as a standard for communicating information on the Web,. Now the XML technology has become a standard format to exchange information over the Internet, and the importance of database technologies that support storage, processing, and delivery of XML is still increasing [4].

### 2.1 DTD

The solution provides a template precast DTD to validate the compliance of a new entry (like files) from the model established. In the literature there are several variants (normalized or not) of the model DTD that are well presented and offers to the user an ergonomic space well done [5].

#### 2.1.1 Example:

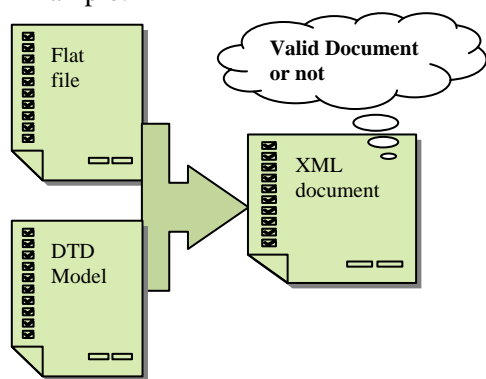


Fig. 1 Transformation from a flat file to an XML file via a model DTD

## 2.2 Schema-Xml and DTD

XML Schema as a recommendation published by the W3C is a language for describing XML document format for defining the structure and content type of an XML document (the syntax). This definition allows in particular to verify the validity of this document.

The XLM Schema is usually used with DTD to validate the documents and together they present a robust [6] tool for transforming a plat file into an XML document. The following figure shows the process.

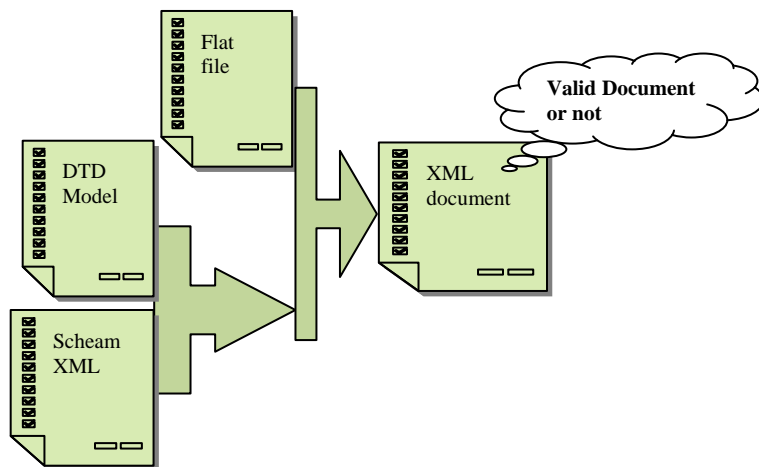


Fig 2 : Transformation from a flat file to an XML file via a model DTD and XML schema.

## 3. Limitation the XML solution

### 3.1 DTD solution

The problem raised in the DTD model is that it has no lexical vision nor semantics of the processed data, so the lexeme read are inserted (in the XML file) by the first come first served without a general understanding of the sequence of tokens. And it may generate additional errors compared to the errors recognized by the OCR. The following examples illustrate the three cases document validation.

#### 3.1.1 Example 1 : Valid document

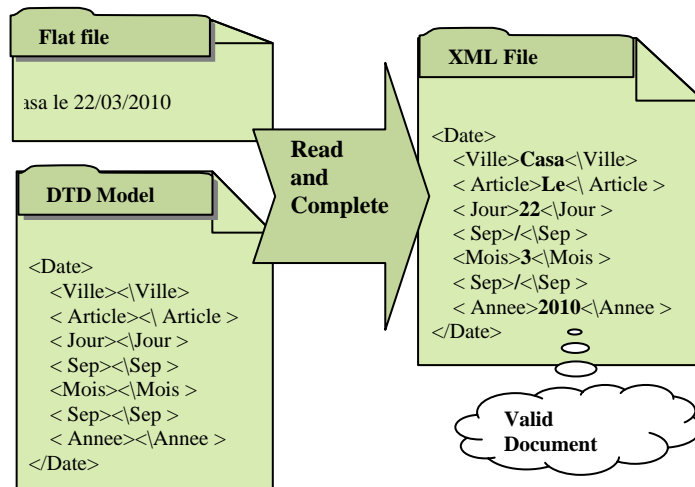


Fig 3: The reading of the document is made successfully, because there were no errors in the flat file.

### 3.1.2 Example 2 : Invalid document

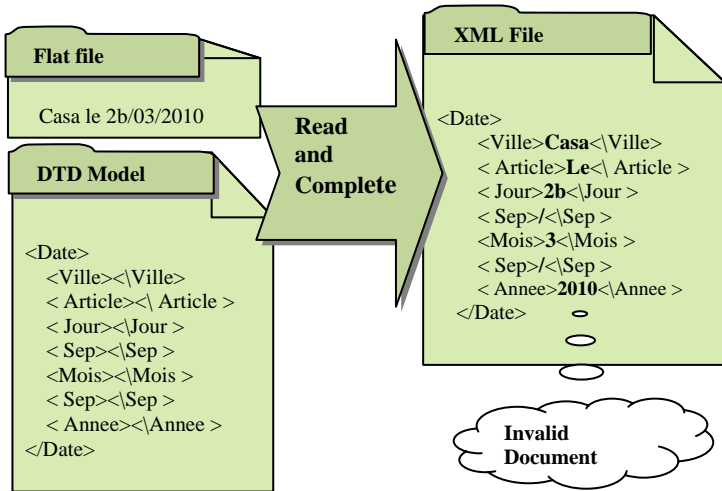


Fig 4: Transformation from a flat file to an XML file via a DTD model with errors recognition.

DTD is not able to detect that the document is invalid, since the meaning of the lexeme (Day 2b) passes unseen.

### 3.1.3 Example 3 : Invalid document

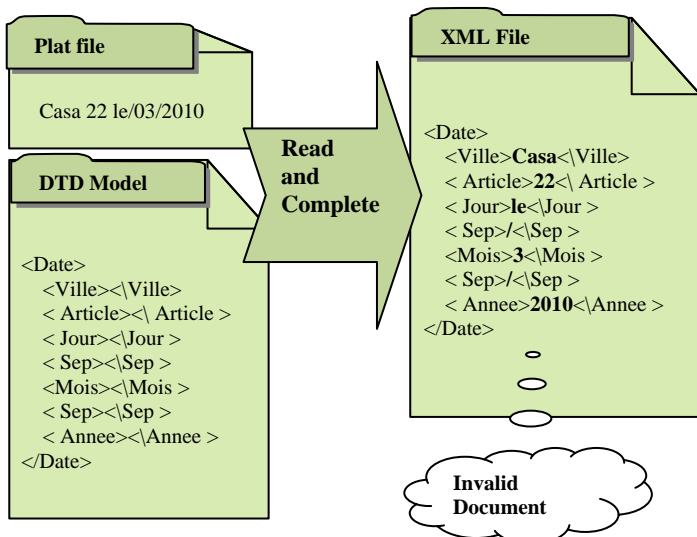


Fig 5: Transformation from a flat file to an XML file via a DTD model with errors recognition.

DDT is still unable to detect grammatical errors, as it displays no recognition error.

### 3.2 DTD + Schema XML solution

XML Schema offers a very good tool for the validation of documents, although it is able to detect lexical and semantic errors, but some errors can escape it because it doesn't operate on the syntactic level. Even Concatenated

with the DTD tool, errors keep showing up in the system. The following examples illustrate the three cases document validation.

### 3.2.1 Example 1: Valid document

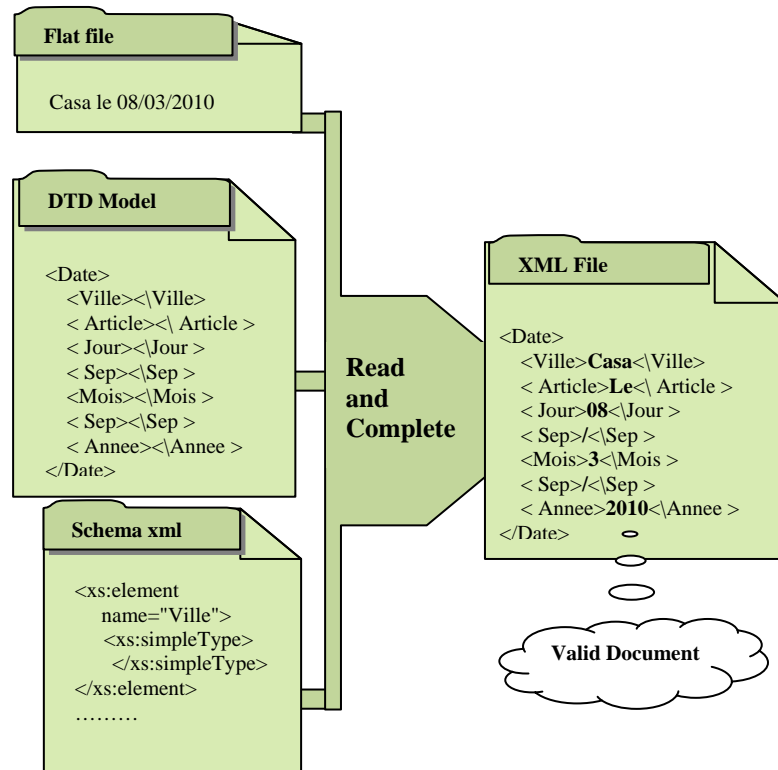


Fig 6: Transformation from a flat file to an XML file via a model DTD and schema xml without recognition errors

The reading of the document is made successfully, because there were no errors in the flat file.

### 3.2.2 Example 2: invalid Document

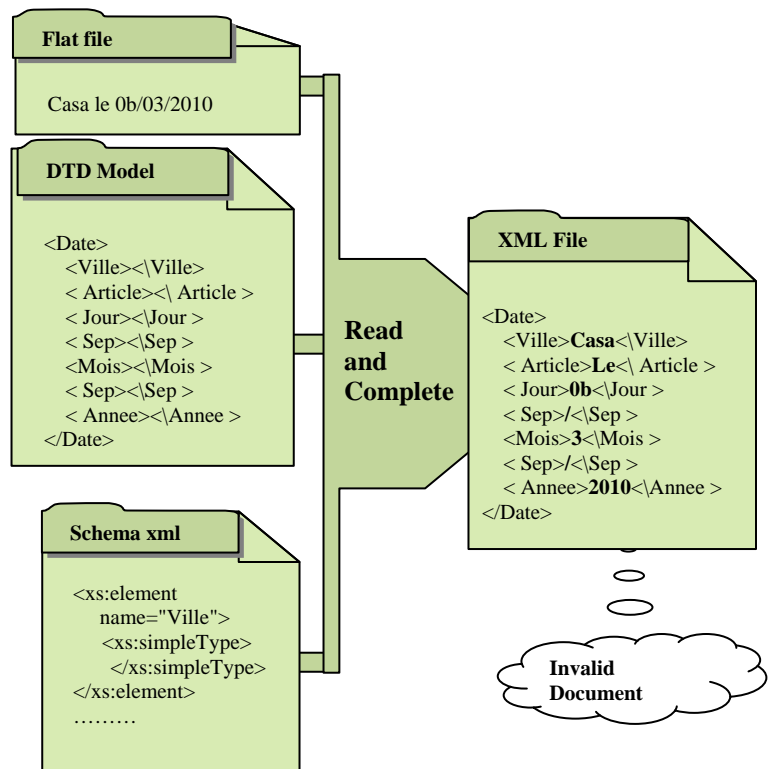




Fig 7: Transformation from a flat file to an XML file via a DTD model and an xml schema with errors recognition.

The xml schema is able to detect the lexical errors (0b :the day field), and also the semantic errors like 23for the month field.

### 3.2.3 Example: invalid Document

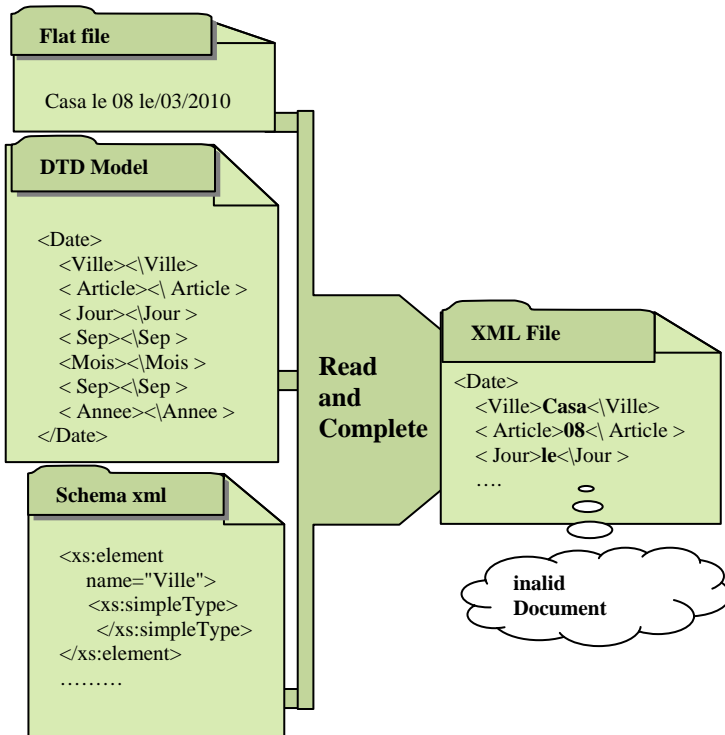


Fig 8: Transformation from a flat file to an XML file via a DTD model and an xml diagram with errors recognition.

The schema xml is not able to detect the syntax errors and they can generate an additional error because they don't have a global vision for the arrival tokens.

### 3.3 Summarization:

The use of xml in the dematerialization of documents is useful and practical, since this technology is capable of detecting a number of mistakes, but it is unable to detect all mistakes, besides, this method is unable to correct errors recognition

This limitation is due to the fact that XML is not a vision of the syntactic outcome studied, and it just treats lexeme by lexeme without a comprehensive framework on the content covered.

## 4. The method suggested

The majority of documents of the invoice type have the same structure [7], and sharing the same zones [8] like: dates zone, customer zone... and many forms such us: prices, tables, logo etc... In other words, it resembles a

structure [7] which belongs to and respects a given language, i.e. grammar with a specific lexicon and syntax.

Our approach is based on the development of a grammar which will give the contents of the documents of the type invoices [9]. Thus, the set up of a grammar per customer will enable us to deduce the structure [10] of its documents and consequently will allow us to detect all the errors related to the grammar of its language and to carry out their correction if possible.

## 5. Grammar of the document

This approach includes three types of analyzers: lexical, syntactic and semantic. Each one of them will be concerned with a specific task which will be described in the following sections. But going further in the description of the types of analyzer, it is initially necessary to define the alphabet on which one will work.

In the literature, an alphabet is a nonempty finite whole of symbols. The latter can be unspecified letters or characters. As well says as a word is in addition to only one sequence of elements of A.(not clear)  
 Practically, the invoice document uses usual symbols like a...z,A...Z..., for that the alphabet [11] adopted for this kind of documents(invoices)doesn't leave this formal framework.

### 5.1 Parser Vs DTD

The syntax of an invoice document can be described by a grammar describing the arrangement of lexical units. The parser receives a sequence of lexical units from the lexical analyzer and must verify if it can be generated by the grammar of language.

#### 5.1.1 Formal definition of grammar [11]

Formally, a grammar is a set denoted as  $G=(V_T, V_N, S, P)$ .

- $V_N$  is a non-empty set of terminal symbols.
- $V_T$  is a non-empty set of non-terminal symbols with  $V_N \cap V_T = \text{empty}$ .
  - S is an initial symbol (axiom).
  - P is a set of production rules.

#### 5.1.2 In pratique

Les us examine the following example:

From a practical standpoint, we consider the following grammar:

Ident = {Casa, Casablanca, Client, Rabat,...}

Chainqqc={a|..|z,A|...|Z,0|1..|9| ;| ?|...}+

Figure=(0|1|2|3|4|5|6|7|8|9)+.

Sep=(-|/|:|.|%).

From the preestablished symbols we can form the following units:

$V_T = \{\text{Ident, Figure, Sep}\}$ .

```

VN={DateZone,ClientZone,QteZone,UnitPriceZone,
VATZone
    ,TTCZone,SumInLettersZone}.
S=Zone.
P={
    Zone→
        DateZone|ClientZone|QteZone|UnitPriceZone|
        TVAZone|TTCZone|SumInLettersZone.
        DateZone → Ident Figure Sep Figure Sep
Figure.
        ClientZone → Ident Sep Chainqqc.
        QteZone → Figure.
        UnitPriceZone → Figure Sep Figure.
        VATZone → figure Sep
        TTCZone → Figure Sep Figure.
        SumInLettersZone → Chainqqc.
    }
    
```

While the DTD is a limited tool because it addresses only the part shape and also some recognition errors could go undetected.

### 5.1.3 Lexical[11] and semantic[12] analyzer Vs schema XML

The principal task of this analyzer is reading the characters of entry and producing as a result a succession of lexemes that the parser will have to treat. Still, it is necessary to define what a lexeme is.

A lexeme is a continuation of characters which has a collective significance. Take this sentence, for example: Casa the 12/03/2009; it can be translated in the following way Ident (Keyword) Ident Chiffre (chiffre or number?!) backslash Chiffre backslash Chiffre. With such an analyzer the detection of a possible lexical error is practically easier and less expensive: both in terms of the memory occupancy rate of the processor, and the complexity of the algorithm used as well.

Generally, a grammar cannot provide a 100% description of the content of a given language, even with the use of two powerful tools: the lexical analyzer and the parser. That is why languages generally rely on a third semantic analyzer [13].

This failure [14] is due to the fact that neither the lexical analyzer nor the parser can detect an error type such as:

- A year estimated at 3000
- A month exceeds 12
- A miscalculation of the TTC.

## 6. Results and discussions

### 6.1 Results

The use of a model based on XML diagram, as well as the model based on our approach gives important results, the latter are given in the form of three fields in particular the average of the existing errors, the average of the detected errors, the average of corrigible errors, and the percentage of correction.

#### 6.1.1 Case 1: without syntax errors

The results obtained during the test are shown in the table below:

Table 1: Average errors

Method used	Number invoices	Average existing Errors	Average detected Errors	Average corrigible Errors	Percentage of correction
Diagram XML	100	35	82	30	85,72%
Our approach	100	33	70	10	30,30%

#### 6.1.2 Case 2: with syntax errors

The results obtained during the test are shown in the table below:

Table 2: Average errors

Method used	Number invoices	Average existing Errors	Average detected Errors	Average corrigible Errors	Percentage of correction
Diagram XML	100	35	82	5	14,29%
Our approach	100	33	63	26	78,79%

#### 6.1.3 Discussion and comparison of the results

The model based on an xml diagram, proposes a multitude of choices concerning the types to be defined, in particular the kind types: string, positive integer... as well as the possibility of generating a regular expression, for that the xml solution is able to read lexeme by lexeme and to test the validity of each chain with share. On the contrary our method treats at the same time lexeme by lexeme as well as the sequence of the continuations of the lexemes according to a given order.

To sum up, the xml solution treats only the lexical part, however our method will operate beyond the lexical part respectively on the syntactic and semantic levels, which partly explains the variation

observed on the level of the rate of the corrigible errors.

A positive account about this method is the ability to offer a correction when there is a syntax error, while this has been impossible through XML. This possibility is offered by the language theory, specially LL<sup>1</sup> and SLR<sup>2</sup> languages [11].

## 7. Conclusion and Implications

In this paper, we proposed a new method based on the theory of languages, it consists in installing a mini compiler which operates via three analyzers: The process of correction which is summarized in the following steps: Reception of the concerned zone, launch of the lexical analyzer, Launch of the parser, launch of the semantic analyzer, then Correction of the errors if possible.

A major advantage of this method is its ability to detect all errors of recognition and most (but not all) of them (all of them or most of them?). Another positive point about this method is that it is about a less expensive solution and especially an easy one to set up.

Our main goal has been to develop a learning tool capable of correcting the errors which are not detected by the mini compiler.

## References

- [1] Rémy Kessler, Juan Manuel, Torres-Moreno et Marc El-Bèze, Classification automatique de courriers électroniques par des méthodes mixtes d'apprentissage - Laboratoire d'Informatique d'Avignon / Université d'Avignon- 2001.
- [2] Yassin Aziz REKIK : Modélisation et manipulation des documents structurés : une approche modulaire, flexible et évolutive. –Thèse N° 2396 (2001)-.
- [3] Cécile Roisin Adaptation aux différents modes de lecture 2004
- [4] E.J.Thomson Fredrick, G.Radhamani: INFORMATION RETRIEVAL USING XQUERY PROCESSING TECHNIQUES International Journal of Database Management Systems ( IJDBMS ), Vol.3, No.1, February 2011.
- [5] Zurinahni Zainol : XML Documents Normalization Using GN-DTD International Journal of Information Retrieval Research, 1(1), 53-76, January-March 2011 53
- [6] Kinsun Tam Sanjay Goel Jagdish S. Gangolly On the design of an XML-Schema based application for business reporting: An XBRL Schema Perspective The International Journal of Digital Accounting Research Vol. 2, No. 1, pp. 83-118.

- [7] Nouredine CHATTI, Sylvie CALABRETTO : MultiX : un formalisme pour l'encodage des documents multi-structurés - LIRIS-INSA de LYON - 2001.
- [8] Rocio Abascal - Michel Beigbeder - Aurélien Bénéel - Sylvie Calabretto -Bertrand Chabbat - Pierre-Antoine Champin - Nouredine Chatti – David Jouve - Yannick Prié - Béatrice Rumpler - Eric Thivant : Modéliser la structuration multiple des documents - LIRIS – INSA de Lyon – 2001.
- [9] Rocio Abascal, Michel Beigbeder, Aurélien Bénéel, Sylvie Calabretto, Bertrand Chabbat, Pierre-Antoine Champin,
- [10] Nouredine Chatti, David Jouve, Yannick Prié, Béatrice Rumpler, Eric Thivant :Documents à structures multiples - LIRIS CNRS FRE-2672 INSA de Lyon 2000-.
- [11] Pierre-Edouard Portier\_, Sylvie Calabretto Modélisation des connaissances dans le cadre de bibliothèques numériques spécialisées - Université De Lyon, INSA de Lyon, LIRIS - 2002.
- [12] Compilation et théorie des langages –Université de Bretagne occidentale-.
- [13] D. Jouve a,b,\*, Y. Amghar a, B. Chabbat b, J.-M. Pinon Conceptual framework for document semantic modelling: an application to document and knowledge management in the legal domain –sciencedirect 29 junaury 2003-.
- [14] David JOUVE : Modélisation sémantique de la réglementation –thèse 03 ISAL 0071 28 novembre 2003-.
- [15] Jean-Luc Minel, Jean-Pierre Desclé, Emmanuel Cartier. Gustavo Crispin, Slim Ben Hazez Agata Jackiewicz :

<sup>1</sup> Left to right scanning, and leftmost derivation.

<sup>2</sup> Shift/Reduce Left right

# Techniques, Advantages and Problems of Agent Based Modeling for Traffic Simulation

Ali Bazghandi

School of computer engineering, Shahrood University of technology  
Shahrood, Semnan, Iran

## Abstract

Agent-based modeling (ABM) is a powerful simulation modeling technique in the last few years.

ABM, as an approach to simulating the behavior of a complex system in which agents interact with each other and with their environment using simple local rules, is gaining popularity and widespread use in many areas. Successes of this approach in predicting traffic flow in metropolitan areas, the spread of infectious diseases, and the behavior of economic systems have generated further interest in this powerful technology.

In this paper we focus on agent-based approach to traffic simulation, and investigate its benefits, difficulties and (microscopic-macroscopic) techniques.

**Keywords:** *ABM, simulation, traffic, benefits, microscopic, macroscopic and problems*

## 1. Introduction

The use of computer systems to solve problems of interest in physics, biology, chemistry, economics, and social sciences has been well-established for decades. The great advances in computing power, software development, computer graphics, communications networks, and a host of other technologies have elevated the domain of applicability to problem solving from simple arithmetic calculations to advanced numerical methods and the creation of large simulations solving myriad systems of complex equations in real-time.

## 2. ABM Advantages

The advantages of ABM over other modeling techniques can be captured in four statements: (i) ABM captures emergent phenomena; (ii) ABM provides a natural description of a system; (iii) ABM is flexible. and (iv) ABM is low cost and time saving approach. It is clear, however, that the ability of ABM to deal with emergent phenomena is what drives the other benefits.

### 2.1 ABM and emergent phenomena Adjacency

Emergent phenomena result from the interactions of individual entities. By definition, they cannot be reduced to the system's parts: the whole is more than the sum of its parts because of the interactions between

the parts. An emergent phenomenon can have properties that are decoupled from the properties of the part. For example, a traffic jam, which results from the behavior of and interactions between individual vehicle drivers, may be moving in the direction opposite that of the cars that cause it. This characteristic of emergent phenomena makes them difficult to understand and predict: emergent phenomena can be counterintuitive. ABM is, by its very nature, the canonical approach to modeling emergent phenomena: in ABM, one models and simulates the behavior of the system's constituent units (the agents) and their interactions, capturing emergence from the bottom up when the simulation is run.

### 2.2 Natural description provided by ABM

In many cases, ABM is most natural for describing and simulating a system composed of "behavioral" entities. Whether one is attempting to describe a traffic jam, the stock market, voters, or how an organization works, ABM makes the model seem closer to reality. For example, it is more natural to describe how vehicles move in a lane than to come up with the equations that govern the dynamics of the density of vehicles. Because the density equations result from the behavior of vehicles, the ABM approach will also enable the user to study aggregate properties.

### 2.3 ABM flexibility

The flexibility of ABM can be observed along multiple dimensions. For example, it is easy to add more agents to an agent-based model. ABM also provides a natural framework for tuning the complexity of the agents: behavior, degree of rationality, ability to learn and evolve, and rules of interactions. Another dimension of flexibility is the ability to change levels of description and aggregation: one can easily play with aggregate agents, subgroups of agents, and single agents, with different levels of description coexisting in a given model. One may want to use ABM when the appropriate level of description or complexity is not known ahead of time and finding it requires some tinkering.

## 2.4 ABM is cost-effective and time saving approach

Champion et. al [1] describe simulation as an effective tool used for reproducing and analysing a broad variety of complex problems, difficult to study by other means that might be too expensive or dangerous. Traffic can be viewed as a complex system [Faieta et. al., Sanford in [2]], therefore simulation is a suitable tool to analyze traffic systems. Traffic simulation is the state-of-the-art method used to assess and evaluate transport schemes for reducing congestion [3]. Rather than implementing a scheme without knowing whether the outcome will be a success, the scheme can be implemented in a simulation to determine its effectiveness:

Infrastructure improvements are costly, hence any such project must be carefully evaluated for its impact on the traffic.[4]

The economic impact of traffic management grows each day. Well-designed and well-managed highway systems reduce the cost of transporting goods, cut energy consumption, and save countless person-hours of driving time. To reduce congestion, many countries have been investing heavily in building roads, as well as in improving their traffic control systems. The 1998 budget for Federal Highway Administration of USA is \$19,680,000,000. Of this amount, \$1,047,000,000 is allocated for congestion mitigation and air quality improvement [US Budget '98] [5].

On the other hand once the computer environment is established for social phenomena, the research cost will be much lower than the traditional research approaches. Also computers can implement complex simulation processes in several minutes at most.

## 3. ABM techniques

Simulations can be classed as continuous or discrete. Continuous models take the form of equations using variables that correspond to real values. By solving the equations, the state of the model at any given point in the simulation can be calculated. Discrete simulations represent reality by modelling the state of the system and its state changes after time or events have passed. There are two types of discrete simulation: discrete time models and discrete event models. Discrete time models (time-sliced) are those that split the simulation into fixed time intervals. At each interval, the state of the model is updated using functions that describe the interactions. Discrete event models (event-oriented) are those which maintain a queue of events scheduled to happen in order of time, each event representing the change of state of an element in the model. The simulator processes the events in order, and each one can alter the event queue. [6], [7], [1]

## 3.1 Macroscopic vs. Microscopic

Traffic simulators can be microscopic or macroscopic depending on the level of detail required. Macroscopic simulators model the flow of traffic using high-level mathematical models often derived from fluid dynamics, thus they are continuous simulations. They treat every vehicle the same, and use input and output variables such as speed, flow and density. These simulators cannot differentiate between individual vehicles, and usually do not cater for different vehicle types. They lack the ability to model complex roadways, detailed traffic control features or different driver behaviours. [4], [8], [9]. Macroscopic simulators are most useful for the simulation of wide-area traffic systems, which do not require detailed modelling, such as motorway networks and interregional road networks [6]. This approach is not very realistic because in real life there are many different types of vehicle driven by different individuals who have their own styles and behaviours. However, it is fast and can be useful and accurate, but is not suited to urban models. [4]

Microscopic simulators model individual entities separately at a high level of detail, and are classed as discrete simulations. Each vehicle is tracked as it interacts with other vehicles and the environment. Interactions are usually governed by car-following and lane-changing logic. Rules and regulations are defined to control what can and cannot be done in the simulation, for example speed limits, rights of way, vehicle speed and acceleration. [6], [9]. Traffic flow details usually associated with macroscopic simulation are the emergent properties of the microscopic simulation. Microscopic simulators can model traffic flow more realistically than macroscopic simulators, due to the extra detail added in modelling vehicles individually [4]. Microscopic simulators are widely used to evaluate new traffic control and management technologies as well as performing analysis of existing traffic operations [9].

A very simple form of microscopic simulation is cellular simulation, which involves modelling the road as a series of cells and moving the vehicles between cells based on vehicle parameters. This method can implement links using an array with length equal to the number of cells in the link. Cell length has to be determined and must be the same for all cells, which is a disadvantage because it assumes all vehicles occupy the same amount of space. When the simulation is run, each cell can be either empty or occupied by one vehicle. Vehicles are moved forwards by their speed and are restricted by vehicles in front. Links are connected to nodes and rules exist which determine where vehicles go when they reach a node. This method can be very efficient because of the simple array structure, but it lacks some realism. [10], [11]. An even simpler approach is queue-based simulation,

where vehicles always move at a set speed until they reach a queue at the end of each link. [12], [13].

What is often mentioned as multi agent based simulation is microscopic modeling of emergent phenomena. Of course, macroscopic parameters can be result of microscopic simulation.

The agents used in the microscopic level for example are following types [14]:

- vehicle agent
- road agent
- intersection agent
- signal agent

### 3.2 Classification of simulation and simulators

Traffic simulations can be broadly classified by the type of road network and features they can simulate. The two main classes for simulators are those designed for motorway and urban environments. Simulators supporting a motorway environment focus on multiple-lane high-speed motorways. Much of the complexity required for a city environment does not need to be modelled, and the simulation can focus on vehicle behaviour and interaction. Motorway environments can be simulated accurately by both macroscopic and microscopic simulators [15]. The main features of a microscopic motorway simulator are car-following and lane-changing behaviours. Junctions are sometimes modelled, allowing entry/exit rate to be varied to test the efficiency of the motorway under varying traffic load. Practical uses include studying the effect of motorway accidents, stop-start congestion, speed limits, ramp metering and lane closures on traffic flow. [16], [15]

An urban environment is one of the most difficult and complex traffic scenarios [4]. In contrast to motorway environments, urban environments have a traffic flow that is interrupted by intersections, traffic lights, roundabouts and other features. In addition to the extra road features, realistic urban simulators should model not only different classes of vehicle, but also pedestrians, cyclists and public transport systems. [17]. Urban traffic networks are usually very complex with many road sections and intersection points, often with conflicting traffic flows [18]. They usually have to manage a large number of vehicles on small road sections, which can result in a large amount of congestion [19]. Microscopic simulators are well suited to urban environments as vehicles can respond individually to the road features. Macroscopic simulators are not able to model the complexity of

urban environments; they are only used to provide abstract flow details.

Some simulators can model both motorway and urban environments at the same time; these are classed as integrated or combined simulators. This is useful for the simulation of large areas encompassing both motorway and urban roads, especially where the performance of one affects the other, and is advantageous to the user as one package can simulate various scenarios. [20], [8]. Some simulators have focussed instead on modelling specific objectives such as to test intelligent vehicle control units, to analyse vehicle safety and comfort, or traffic at toll booths [7].

## 4. Problems & Challenges

Considering the nature of social phenomena with too many (known & unknown) complex factors is the first problem in simulating these systems (For example in traffic systems, driver behaviors vary dramatically with geographic location and change over time. In reality they most often involve human agents, with potentially irrationally behavior, subjective choices, and complex psychology—in other words, soft factors, difficult to quantify, calibrate, and sometimes justify). Of course, it is better that we count this problem as a characteristic of social phenomena.

Although a lot of academic attention has been given to the subject, there are very few traffic applications, perhaps because of the “soft” nature of the variables and the difficulty in measuring parameters. Social simulation in traffic has not been very successful so far, because the emphasis has been on using it as a predictive tool rather than as a learning tool. For example, a traffic engineer can understand congestion better by playing with an agent-based model of it. Then, of course, quantifying the tangible benefits of something intangible is difficult, and a traffic engineer cannot claim to have reduced congestion of a lane by playing with a simulation of vehicles.

One issue related to the application of ABM to the traffic simulation is common to all modeling techniques: a model has to serve a purpose; a general-purpose model cannot work. The model has to be built at the right level of description, with just the right amount of detail to serve its purpose.

Another problem is that a lot of agent based toolkits include performance limitations: with a large number of agents, execution speed drops considerably. Usually these tools are not designed for extensive simulations.

## 5. Conclusions

Although above problems may constitute a major source of problems in interpreting the outcomes of

simulations, it is fair to say that in most cases ABM is simply the only game in town to deal with such situations. Having said that, one must be careful, then, in how one uses ABM: for example, one must not make decisions on the basis of the quantitative outcome of a simulation that should be interpreted purely at the qualitative level. Because of the varying degree of accuracy and completeness in the input to the model (data, expertise, etc.), the nature of the output is similarly varied, ranging from purely qualitative insights all the way to quantitative results usable for decision-making and implementation.

The last major issue in ABM is a practical issue that must not be overlooked. By definition, ABM looks at a system not at the aggregate level but at the level of its constituent units. Although the aggregate level could perhaps be described with just a few equations of motion, the lower-level description involves describing the individual behavior of potentially many constituent units. Simulating the behavior of all of the units can be extremely computation intensive and therefore time consuming. Although computing power is still increasing at an impressive pace, the high computational requirements of ABM remain a problem when it comes to modeling large systems.

## References

- [1] CHAMPION, A. et al., 1999. Traffic generation with the SCANer II simulator: towards a multi-agent architecture. DSC '99: Proceedings of the first Driving Simulation Conference, 1999, pp 311-324
- [2] EROL, K. et al., 1998. Application of Agent Technology to Traffic Simulation. [online] Available from: <http://www.tfhr.gov/advanc/agent.htm> [Accessed April 2006]
- [3] CLARK, J. and DAIGLE, G., 1997. The importance of simulation techniques in ITS research and analysis, WSC '97: Proceedings of the 29th conference on Winter simulation, 1997, ACM Press pp1236-1243.
- [4] EHLERT, P. and ROTHKRANTZ, L., 2001. A Reactive Driving Agent for Microscopic Traffic Simulation, ESM '01: Proceedings of the 15th European Simulation Multiconference, 2001, SCS Publishing house pp943-949
- [5] Kutluhan Erol., Renato Levy. and James Wentworth. 1999. Application of Agent Technology to Traffic Simulation, United States Department of Transportation - Federal Highway Administration ([www.tfhr.gov/advanc/agent.htm](http://www.tfhr.gov/advanc/agent.htm))
- [6] SCHULZE, T. and FLIESS, T., 1997. Urban traffic simulation with psycho-physical vehicle-following models, WSC '97: Proceedings of the 29th conference on Winter simulation, 1997, ACM Press pp1222-1229.
- [7] ALGERS, S. et al., 1997. Review of Micro-Simulation Models. SMARTEST Deliverable D3, Institute for Transportation Studies, University of Leeds, Leeds, UK, Aug. 1997.
- [8] BOXILL, S. A. and YU, L., 2000. An Evaluation of Traffic Simulation Models for Supporting ITS Development. Center for Transportation Training and Research, Texas Southern University, October 2000.
- [9] OWEN, L.E. et al., 2000. Street and traffic simulation: traffic flow simulation using CORSIM, WSC '00: Proceedings of the 32nd conference on Winter simulation, 2000, Society for Computer Simulation International pp1143-1147.
- [10] TANG, W. and WAN, T.R., 2005. Multi-agent Animation Techniques for Traffic Simulation in Urban Environment. WSCG (Short Papers), 2005, pp161-164.
- [11] NAGEL, K., Draft as of 02 Feb 2004. Multi-agent Transportation Simulation. [online] Available from: <http://www.vsp.tu-berlin.de/publications/matsim-book/> [Accessed December 2005]
- [12] NAGEL, K., and SCHRECKENBERG, M., 1992. A Cellular Automaton Model for Freeway Traffic. Journal de Physique I, 2, pp 2221-2229.
- [13] NAGEL, K., 2004. Multi-Agent Traffic Simulations. Presented at ETH Zurich: Institute for Computer Science.
- [14] Kosuke Ono<sup>1</sup>, Takashi Hata<sup>1</sup>, Toyofumi Maetani<sup>1</sup>, Masateru Harao and Kouichi Hirata<sup>2</sup>, T. Washio et al. "Development of a Multi-Agent Based Generic Traffic Simulator". (Eds.): JSAI 2005 Workshops, LNAI 4012, pp. 249-260, 2006. c\_Springer-Verlag Berlin Heidelberg 2006.
- [15] RIGOLLI, M. and BRADY, M., 2005. Towards a behavioural traffic monitoring system, AAMAS '05: Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems, 2005, ACM Press pp449-454.
- [16] CHAMPION, A. et al., 1998. Behavioral Road Traffic Simulation with ARCHISIM. SCSC: Proceedings from the Summer Computer Simulation Conference, Society for Computer Simulation International, pp. 359-364.
- [17] DONIKIAN, S. and PHANE, 2001. HPTS: A Behaviour Modelling Language For Autonomous Agents, AGENTS '01: Proceedings of the fifth international conference on Autonomous agents, 2001, ACM Press pp401-408.
- [18] LINDSEY, R. and VERHOEF, E., 2002. Congestion Modelling. Handbook of Transport Modelling. 1st reprint 2002 edn. Pergamon, pp. 377-397.
- [19] TOMAS, V.R. and GARCIA, L.A., 2005. A cooperative multiagent system for traffic management and control, AAMAS '05: Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems, 2005, ACM Press pp52-59.

- [20] PREVEDOUROS, P.D. and WANG, Y., 1999.  
Simulation of a Large Freeway/Arterial Network  
with CORSIM, INTEGRATION and WATSim.  
Paper presented at the 78th Transportation  
Research Board Annual Meetings, Washington,  
DC. January 1999.



# Optimization Parameters of tool life Model Using the Taguchi Approach and Response Surface Methodology

Kompan Chomsamutr<sup>1</sup>, Somkiat Jongprasithporn<sup>2</sup>

<sup>1,2</sup>Department of Industrial Engineering, King Mongkut's University of Technology North Bangkok, Thailand

## Abstract

The objective of this research is to compare the cutting parameters of turning operation the work pieces of medium carbon steel (AISI 1045) by finding the longest tool life by Taguchi methods and Response Surface Methodology: RSM. This research is to test the collecting data by Taguchi method. The analyses of the impact among the factors are the depth of cut, cutting speed and feed rate. This research found that the most suitable response value; and tool life methods give the same suitable values, i.e. feed rate at 0.10 mm/rev, cutting speed at 150 m/min, and depth of cut at 0.5 mm, which is the value of longest tool life at 670.170 min, while the average error is by RSM at the percentage of 0.07 as relative to the testing value.

Keywords: *Tool life, Taguchi Method, Response Surface Methodology, Cutting parameters*

## 1. Introduction

The Computer Numerical Control(CNC) machine refers to the automation of machine tools to produce the good quality of work pieces. The production of low cost and good quality of work pieces compose of various factors. The research study found that the parameters are the main factors of lathing to get the job done. Meanwhile it is the factor of fixing tool life. If the selection of parameters is not correct, it will cause the shorter tool life. This will cause the impact of cost of production which could not compete to market effectively. The research found that the Taguchi method and Response Surface Methodology are the suitable parameters.

Response Surface Methodology(RSM) is a statistical and mathematical method that is suitable for development and improving the process, and finding the best value for the process, which in the mean time, test design such as Fractional Factorial Design, Central Composite Design (CCD), Box-Behnken Design(BBD), and orthogonal arrays are designed to use few tests to find relationship of each factor. While Taguchi method is a method that cannot be used to analyze Interaction Effect for every case without any consideration to Confounding of the factor, it

does not use sampling principle of test design due to economical test cost requirement. Efficiency of this method depends on the selection of suitable orthogonal array. However, Taguchi method is still popular for engineering work, especially in manufacturing sector.

Noorul and Jeyapaul [1] adopted orthogonal array, Grey relational analysis in the ANOVA using Taguchi method to find suitable level of indentified parameters, and significant association of parameters in order to increase multiple response efficiency of parameters in driller operation for Al/SiC. Later, Mohan et al. [2] adopted Design of Experiments, ANOVA in measuring the data from the collection and analyzed the result with software package MINITAB14 with the objective to increase efficiency of parameters of drilling process of Glass-fiber Polyester material to obtain good surface roughness and low cutting thrust, torque. Then, Kilickap [3] adopted ANOVA, analysis of signal-to-noise ratio to find suitable parameters for the cutting based on Taguchi method. Palanikumar et al. [4] adopted Response Surface Methodology in analyzing the variance for verification model in order to increase efficiency of cutting parameters for surface roughness in the operation of PCD cutting machine with Al/SiC material. Tosum and Ozler [5] used the Taguchi method to investigate multiple performance characteristics and the improvement of optimal cutting parameters in hot turning operations. Thanizhmanil et al. [6] proposed the efficient use Taguchi's parameter design to obtain optimum condition because it leads to minimum number by experimental and lower cost. Rossella *et al.* [7] proposed a new optimization method of the manufacturing parameters using Taguchi method. They found that the experiment design of the orthogonal array of the Taguchi method can identify the significant foaming parameters the adjusted the process.

The above researches showed that Taguchi method and Response Surface Methodology are techniques that increase efficiency successfully applied in industrial work for the best option of the process parameters in the area of machinery. Taguchi method has potential for saving test time and cost relating to the product or manufacturing process development, and quality improvement.

Therefore the study of behavior and relationship of parameters are the main factors namely depth of cut,

cutting speed and feed rate. These parameters can determine the CNC program to modify easily. The researcher has chose such parameters to collect and analyze the information as the form of Taguchi method. The appropriate value of tool life is calculated by Response Surface Methodology.

## 2. Experimental procedure

### 2.1 Test Specimen

This research conducted a cutting test with an automatic machine called PINACHO, RAYO 180 model, using an Insert of KENNAMETAL-KC5010 with Nose Radius of 0.4 millimeter, and SHELLDROMUS OIL B for heat ventilation. The material used in the cutting is medium carbon steel(AISI 1045).

### 2.2 Tool life measurement

The cutting for calculating the working life of tool life is shown as the table of orthogonal array  $L_9$  by cutting the part at 10 rounds of one piece to get the cutting length at 1,000 millimeter and stopping watch and checking the wearing out by the Toolmaker's Microscope of TM 505 which the lenses of 30 times and measure every 1,000 millimeter until the size of the Flank Were are more than 0.6 millimeter(VB max>0.6 millimeter)

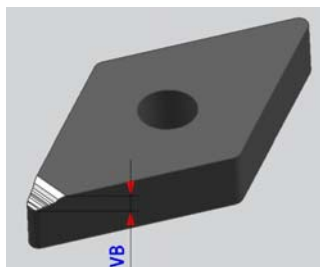


Fig 1: Show machine used to measured of tool life and Flank Were (VB)

### 2.3 Taguchi Method and design of experiment

The Taguchi method is a quality tool that helps improve the work efficiently. It is possible to select suitable factors as shown in Table 1, which indicates factors and their levels in the cutting experiment with CNC machine, which contains 3 factors, and each factor has 3 levels. The table 2 is shown the form of orthogonal array  $L_9$  for data collection. The researcher collects data by 9 conditions. Each condition will be determined by the factors for instance the the first condition is identified by the depth of cut at 0.5 mm., cutting speed at 150 m/min and feed rate at 0.10 mm/rev as well. Therefore, the orthogonal array is selected with  $L_9(3^3)$  method which gave the order of experiment as shown in Table 3.

As mentioned earlier, Taguchi method is used for tuning the turning process by optimizing the process parameters for best tool life. In general, the parameter optimization process of the Taguchi method is based on 8-steps of planning, conducting and evaluating results of matrix experiments to determine the best levels of control parameters [8]. Those eight steps are given as follows.

- Identify the performance characteristics (responses) to optimize and process parameters to control (test).
- Determine the number of levels for each of the tested parameters.
- Select an appropriate orthogonal array, and assign each tested parameters into the array.
- Conduct an experiment randomly based on the arrangement of the orthogonal array.
- Calculate the  $S/N$  ratio for each combination of the tested parameters.
- Analysis the experimental result using the  $S/N$  ratio and ANOVA test.
- Find the optimal level for each of the process parameters.
- Conduct the confirmation experiment to verify the optimal process parameters.

Table 1: Cutting parameters and their levels

Factors	Cutting Parameters	Levels			unit
		1	2	3	
1	Depth of cut (D)	0.50	1.00	1.50	mm
2	Cutting speed (Vc)	150	200	250	m/min
3	Feed rate (F)	0.10	0.15	0.20	mm/rev

Table 2: Table of Taguchi designs (Orthogonal Arrays L<sub>9</sub>)

Experiment Number	Cutting parameter level		
	A	B	C
	Depth of cut	Cutting speed	Feed rate
1	1 (Level 1)	1 (Level 1)	1 (Level 1)
2	1 (Level 1)	2 (Level 2)	2 (Level 2)
3	1 (Level 1)	3 (Level 3)	3 (Level 3)
4	2 (Level 2)	1 (Level 1)	2 (Level 2)
5	2 (Level 2)	2 (Level 2)	3 (Level 3)
6	2 (Level 2)	3 (Level 3)	1 (Level 1)
7	3 (Level 3)	1 (Level 1)	3 (Level 3)
8	3 (Level 3)	2 (Level 2)	1 (Level 1)
9	3 (Level 3)	3 (Level 3)	2 (Level 2)

Table 3: Experimental layout based on an L<sub>9</sub> orthogonal array

Experiment number	Cutting Parameter Level			Parameter setting
	A	B	C	
	Depth of cut	Cutting speed	Feed rate	
1	0.5 (Level 1)	150 (Level 1)	0.10 (Level 1)	A1B1C1
2	0.5 (Level 1)	200 (Level 2)	0.15 (Level 2)	A1B2C2
3	0.5 (Level 1)	250 (Level 3)	0.20 (Level 3)	A1B3C3
4	1.0 (Level 2)	150 (Level 1)	0.15 (Level 2)	A2B1C2
5	1.0 (Level 2)	200 (Level 2)	0.20 (Level 3)	A2B2C3
6	1.0 (Level 2)	250 (Level 3)	0.10 (Level 1)	A2B3C1
7	1.5 (Level 3)	150 (Level 1)	0.20 (Level 3)	A3B1C3
8	1.5 (Level 3)	200 (Level 2)	0.10 (Level 1)	A3B2C1
9	1.5 (Level 3)	250 (Level 3)	0.15 (Level 2)	A3B3C2

Table 4: L<sub>9</sub>(3<sup>3</sup>) Orthogonal Array, Experiment results and S/N ratio

Experiment Number	Cutting Parameters			Measure Tool Life (Min)	S/N ratio
	Depth of cut	Cutting speed	Feed rate		
1	0.5	150	0.10	670.17	56.5237
2	0.5	200	0.15	308.10	49.7738
3	0.5	250	0.20	114.75	41.1951
4	1.0	150	0.15	253.46	48.0782
5	1.0	200	0.20	182.00	45.2014
6	1.0	250	0.10	176.85	44.9521
7	1.5	150	0.20	124.65	41.9138
8	1.5	200	0.10	239.85	47.5988
9	1.5	250	0.15	90.77	39.1588

As the reason of decreasing the numbers of testing suit to the condition by the consideration of the ratio of impact (S/N ration). The analysis of impact between the factors of Higher is better. Type problem for tool life(eq.1) and the analysis of variance(ANOVA) to test the difference of factor levels.

Higher is Better Type Problem for Tool Life

$$S/N = -10 \log \sum_{i=1}^n \frac{1/y_i^2}{n} \quad (1)$$

## 2.4 Response Surface Methodology: RSM

Response surface methodology:(RSM) is usually considered in the context of experimental design as a statistical method for modeling and analyzing of problems in which different variables affect a response of interest. The first step in RSM is to determine a suitable approximation for the actual functional relationship between the response variable y and a set of independent variables as follows; [9].

$$y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \sum_{i=1}^k \beta_{ii} x_i^2 + \sum_{i < j} \beta_{ij} x_i x_j + \epsilon \quad (2)$$

Where  $\beta$  are the coefficients which are calculated using an appropriate method such as the least square method. When resulted estimated surface is an adequate approximation of the true response function, the results will be approximately equivalent to analysis of the actual system. The model parameters can be approximated whenever proper experimental designs are used to collect the data.

## 3. Determination of cutting parameters

Test result collection according to the order of the experiment in Table 4, and analyze S/N ratio of tool life is as follows;

This section provides the results of S/N ratio, main effect plot and ANOVA. From the results of mean S/N ratio and ANOVA analysis, the optimal combination of cutting parameters is achieved and verification tests have been performed to predict the improvement.

### 3.1 Analysis of the signal-to-noise(S/N) ratio

In Taguchi method, the term signal represents the desirable value, and noise represents the undesirable value. Process parameters with the highest S/N ratio always give the best quality with minimum variance [10]. The S/N ratio for each parameter level is calculated by finding the average of S/N ratios at the corresponding level. Fig 2 shows the response table for S/N ratio of tool life for larger is better obtained for different parameter levels.

Response Table for Signal to Noise Ratios of Tool Life Larger is better			
Level	D	Vc	F
1	49.16	48.84	49.69
2	46.08	47.52	45.67
3	42.89	41.77	42.77
Delta	6.27	7.07	6.92
Rank	3	1	2

Fig.2 Response table for Signal to Noise Ratios of tool life

The analysis of S/N ratio of tool life found that the first factor that causes tool life to be great is cutting speed, having feed rate and depth of cut as secondary factors, respectively. After that, the analysis is made to determine suitable factor of each main factor from S/N ratio as shown in Fig.3.

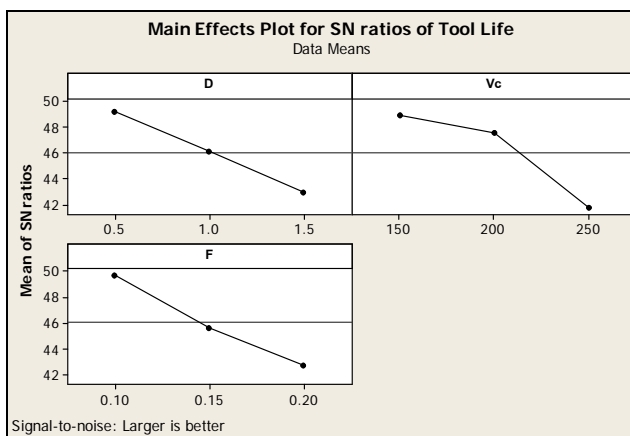


Fig.3 Main effects plot for Signal to Noise Ratios of tool life

Analysis of Variance for SN ratios of Tool Life							
Source	DF	Seq SS	Adj SS	Adj MS	F	P	
D	2	59.044	59.044	29.522	15.55	0.060	
<b>Vc</b>	<b>2</b>	<b>84.842</b>	<b>84.842</b>	<b>42.421</b>	<b>22.35</b>	<b>0.043</b>	<b>Significant</b>
F	2	72.488	72.488	36.244	19.10	0.050	
Residual Error	2	3.796	3.796	1.898			
Total	8	220.169					

Fig.4 Results of ANOVA using data from Signal to Noise Ratios for tool life

Fig.4 shows the results of ANOVA for tool life for a level of significance of 5%(0.05). From ANOVA Fig.4, it is found that, Vc(Cutting Speed) is the significant parameter on tool life. The depth of cut and feed rate is found to be insignificant from ANOVA for tool life study.

In the impact analysis of S/N ratio, factor level will be selected to give maximum S/N ratio as the most suitable factor level. The selection of the most suitable factor level from the graph found that level of factor that causes tool life to be great is when level of depth of cut is 0.5 mm, level of cutting speed is 150 m/min and level of feed rate is 0.10 mm/rev.

### 3.2 Analysis of variance

The ANOVA study performed to investigate the statistical significance of the process parameters affecting the response(tool life). This is achieved by separating the total variability of the S/N ratios, which is measured by the sum of the squared deviations from the total mean of the S/N ratio, into contributions by each of the process parameters and the error [11]. F-test is carried out to judge the significant parameter affecting the tool life. The larger F-value affects more on the performance characteristics.

## 4. Development of Response Surface Model

The analysis with Taguchi method mentioned above is an analysis only for the main factors that affect tool life without any consideration of correlation between factors.

Therefore, the researcher has performed Response Surface Regression in the analysis of correlation between factors. Analysis with response optimizer function that finds the best value for the third factor at the significant level of 95% by response analysis as follows;

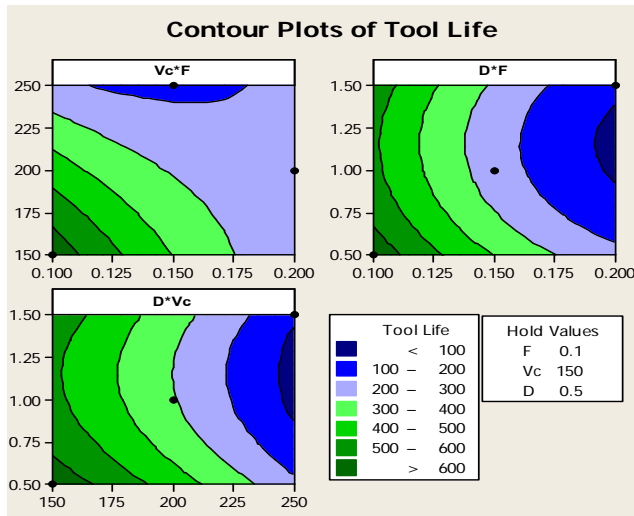


Fig.5 Contour Plots of tool life

The above analysis found that contour plots of tool life is a curve. Therefore, mathematical model suitable for predicting suitable value is Quadratics model by considering the Full Quadratics model as shown in equation 3, which the coefficients of factors that affect response value are as shown in Table 5.

Table 5: Coefficients of Factors that affect Response Value

Term	Coef of tool life
Constant	2800.51000
D	-1536.13000
Vc	-5.83803
F	-11866.63333
D*D	215.78000
Vc*Vc	-0.00205
F*F	35796.66667
D*Vc	4.44453
D*F	20.53333

With the above coefficients of factors that affect response value above, a mathematical model equation can be built as follows;

Mathematical model for forecasting tool life

$$\begin{aligned}
 \text{Tool Life} = & 2800.51 - 1536.13D - 5.83803Vc - \\
 & 11866.63333F + 215.78D^2 - 0.00205Vc^2 - \\
 & 35796.66667F^2 + 4.44453DVc + \\
 & 20.53333DF
 \end{aligned} \tag{3}$$

The model of the appropriate parameters of tool life as the 3rd equation is the comparison between the real value and the forecasting value of the model by the parameter as shown in Table 4 as follows:

Table 6: Comparison of actual value and forecasting value of tool life

Experiment Number	Cutting parameters			Tool life		
	Depth of cut	Cutting speed	Feed rate	Actual	Forecasting	%Error
1	0.5	150	0.10	670.17	670.230	0.01%
2	0.5	200	0.15	308.10	308.207	0.03%
3	0.5	250	0.20	114.75	114.917	0.15%
4	1.0	150	0.15	253.46	253.520	0.02%
5	1.0	200	0.20	182.00	182.107	0.06%
6	1.0	250	0.10	176.85	177.017	0.09%
7	1.5	150	0.20	124.65	124.710	0.05%
8	1.5	200	0.10	239.85	239.957	0.04%
9	1.5	250	0.15	90.77	90.937	0.18%
Average						0.07%

The information in Table 6 shows the result from the comparison between actual value and forecasting value which found that the forecasting values of tool life has the average error of only 0.07%.

The T-Test for the analysis of data difference from the assumption that the average the actual value is not equal to the average forecasting value at the significant level of 0.01 found that P-Value is 0.0001. Therefore, it is concluded that the main assumption should be rejected and the secondary assumption should be accepted, which the average actual value is equal to the average forecasting value.

For finding suitable point of the factors, which is the best point for this experiment using Minitab Release 15, Response Optimizer function, the researcher selected Desirability Function to find suitable value of the factors. After the assessment, the obtained values are as follows;

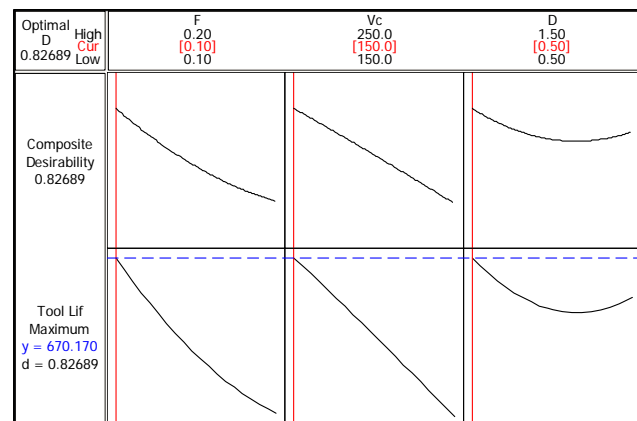


Fig.6 the appropriate value of each factor which effect to tool life

The appropriate value of tool life is the maximum and longest value at 670.170 min. The determination of depth of cut as 0.5 mm, cutting speed as 150 m/min, and feed

rate as 0.10 mm.rev as the satisfaction as the suitable value at 0.82689 as well.

## 5. Conclusion

As the parameter testing of lathing work pieces as depth of cut, cutting speed and feed rate as the surface response of tool life by Taguchi method and Response Surface Methodology as both appropriate value of both methods as shown in Table 7

Table 7: The comparison of cutting parameters by Taguchi method and RSM

Symbol	Cutting Parameters	Tool Life(min)		Unit
		Taguchi	RSM	
D	Depth of cut	0.5	0.5	mm.
Vc	Cutting speed	150	150	m/min
F	Feed rate	0.10	0.10	mm/rev
		670.170	670.230	

As can be seen from Table 7 it found that the suitable of response of tool life by both methods will get the suitable values namely depth of cut at 0.5 mm. cutting speed at 150 m/min and feed rate at 0.10 mm/rev. All mentioned values cause the longest tool life at 670.170 min by Taguchi method and 670.230 by RSM respectively.

## 6. References

[1] Noorul H.A., Marimuthu P. and Jeyapaul R., 2007. Multi response optimization of machining parameters of drilling Al/Sic metal matrix composite using grey relational analysis in the Taguchi method. *Int.J Adv Manuf Technol*: 250-255.

[2] Mohan S., Venugopal A., Rajadurai and Mannan S.L., 2008. Optimization of the Machinability of the Al-SiC Metal Matrix Composite Using the Dynamic Material Model. *Int. Metallurgical and Materials Transaction*. 39 : 2931-2940.

[3] Kilickap E., 2010. Optimization of cutting parameters on elimination based on Taguchi method during drilling of GFRP composite. *Expert Syst. Appl*. 37 : 6116-6122.

[4] Palanikumar K., Shanmugam K., Paulo D.J., 2010. Analysis and optimisation of cutting parameters for surface roughness in machining Al/SiC particulate composites by PCD tool. *International Journal of Materials and Product Technology*. 37 : 117-128.

[5] Tosun N. and Ozler L., 2004. Optimization for hot turning operations with multiple performance characteristics. *International Journal Advance Manufacturing Technology*. 23 : 777-782.

[6] Thamizhmanil S., Saparudin S. and Hasan S., 2007. Analyses of surface roughness by turning process using Taguchi method. *Journal of Achievements in Materials and Manufacturing Engineering*. 20 : 503-505.

[7] S. Rossella, A. Luigi, D. Antonio and B. Giancarlo. Application of Taguchi method for the multi-objective optimization of aluminum foam manufacturing parameters, *International Journal Material Forming* (2009).

[8] T.R. Lin. Experimental design and performance analysis of TiN-coated carbide tool in face milling stainless steel, *journal of Materials processing Technology* 127 (2002), pp. 1-7.

[9] Montgomery D.C., 1991. Design and analysis of experiments. John Wiley and Sons, New York.

[10] Phadke M.S., 1989. Quality engineering using robust design. Englewood Cliffs, NJ: Prentice-Hall.

[11] Nalbant M. et al., 2007. Application of Taguchi method in the optimization of cutting parameters for surface roughness in turning. *Materials and Design*. 28 : 1379-1385.

# Temperature effects on the Drain Current in GaN Dual-Gate MESFET using Two-Dimensional Device Simulation

Hamida DJELTI<sup>1</sup>, Mohammed FEHAM<sup>1</sup>, Achour OUSLIMANI<sup>2</sup> and Abed-Elhak KASBARI<sup>2</sup>

<sup>1</sup> Laboratoire des Systèmes et Technologies de l'Information et de la Communication (STIC)  
Département de Génie Electrique, Faculté de Technologie  
Université Abou Bekr Belkaid de Tlemcen  
BP 230, Chetouane, 13000 Tlemcen - Algérie

<sup>2</sup> ENSEA, Electronique et Commande de Systèmes ECS-Lab EA3649  
6 avenue du Ponceau, 95014 Cergy Cedex, France

## Abstract

Temperature dependence of the GaN-Dual-Gate MESFET (GaN-DGMESFET) DC-characteristics is investigated using two dimensional numerical simulations. Differential equations derived from a Hydrodynamic electron transport model describe the physical proprieties of the device. Simulation results over a wide range of temperature from 300 K to 900 K performed on an industrial software Atlas from SILVACO are presented for a GaN-DGMESFET with a gate length of 0.5  $\mu\text{m}$ . The results show a significant degradation of the DC characteristics. Variation of the electron temperature with the drain-source voltage ( $V_{ds}$ ) is studied and a large temperature is observed for  $V_{ds} > 1$  V. At low drain-source voltage ( $V_{ds} < 1$  V) the electron temperature is closed to the lattice temperature.

**Keywords:** DGMESFET, GaN, Temperature, Steady-state.

## 1. Introduction

In recent years, GaN-based field effect transistors (FETs) have emerged as a promising candidate for high power, high temperature microwave applications and power electronics. These devices' impressive performance is due to the material's properties, such as wide band gap, high breakdown field, and high electron saturation velocity [1, 2] and relatively high electron mobility. Because of the relatively higher band gap energy of GaN (3.47 eV at 300 K) the onset of diffusion-dominated leakage currents generally occurs at much higher temperatures than in GaAs. This provides a potential advantage for GaN IC's. DGMESFET have been commonly used at very high frequency in many different applications such as gain controlled amplifiers, frequency multipliers, phase shifters, stabilised oscillators, power combiners and splitters, [3].

In general, a DGMESFET is basically modelled as a cascade connection of two single-gate MESFET's (SGMESFET) FET<sub>1</sub> and FET<sub>2</sub> as shown in figure 1, where each FET part has a current generator [4]. This configuration improves the output impedance, reduces the feedback capacitance and features a reduction of short-channel effects compared to those observed in single-gate FETs [5].

The temperature effect on the DC characteristics of GaAs DGMESFET for both planar and vertical structures over a wide range of temperature from 250 K to 400 K has been reported [6, 7, 8]. The temperature effects on the DC GaN-DGMESFET characteristics have received a little attention. Indeed, much of the working on the GaN MESFET devices has concentrated their effort on the DC, AC and noise [9, 10].

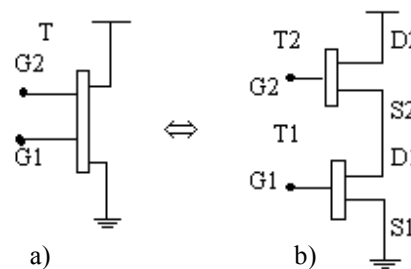


Fig. 1 Symbolic diagram of: (a) DGMESFET, (b) Cascade circuit of two SGMESFET's [4].

In this paper, we investigate the temperature dependence of the DC characteristics in 0.5  $\mu\text{m}$  gate length GaN DGMEFET. The industrial software Atlas from SILVACO, is used to perform the two-dimensional numerical simulations. The transport properties of the GaN DGMEFET are described via differential equations derived from the Hydrodynamic electron transport model (HM). This model consists of an additional coupling of the current density to the carrier temperature.

In the following, section 2 describes the theoretical study of the temperature dependence of the DGMEFET drain current. The results and discussions of a 0.5  $\mu\text{m}$  DGMEFET with a source-to-drain spacing of 3  $\mu\text{m}$  are presented in section 3. Finally, the conclusion is presented in section 4.

## 2. Theoretical analysis of GaN DGMEFET

The most important main factors responsible for the change of DGMEFET performances with the temperature are: the energy band gap, the electron mobility, the saturation velocity and the threshold voltage. The band gap energy is modeled using [11]:

$$E_{gap}(T) = E_{gap}(T = 300\text{K}) + \alpha \left[ \frac{300^2}{300 + \beta} - \frac{T_L^2}{T_L + \beta} \right] \quad (1)$$

Where the values of  $\alpha$  and  $\beta$  for GaN are respectively  $0.909 \cdot 10^{-3} \text{ eV/K}$  and  $774 \text{ K}$ .  $T_L$  is the lattice temperature. Figure 2 shows the GaN band gap energy which decreases with increased temperature.

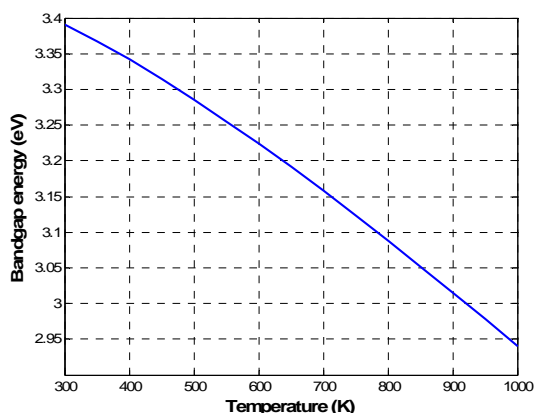


Fig. 2 Temperature dependence of energy band gap.

The empirical expression for the electron mobility suggested by Farahmand [12] is given by (2):

$$\mu_n(T, N) = \mu_{1n} \left( \frac{T_L}{300} \right)^\beta + \frac{(\mu_{2n} - \mu_{1n}) \left( \frac{T_L}{300} \right)^\Delta}{1 + \left( \frac{N}{N_{crit} \left( \frac{T_L}{300} \right)^\gamma} \right)^\epsilon} \quad (2)$$

Where the mobilities at low electric field  $\mu_{1n}$  and  $\mu_{2n}$  are respectively equal to  $295$  and  $1460.7 \text{ cm}^2/\text{V.s}$ , the quoted value of  $\alpha, \beta, \gamma, \Delta$  and  $\epsilon$  are respectively equal to  $0.66, -1.02, -3.84, 3.02, 0.81$ .  $N$  is the local impurity concentration and  $N_{crit}$  is set to  $10^{17} \text{ cm}^{-3}$ .

Figure 3 shows the temperature effect on the electron mobility for three values of doping concentration ( $2 \times 10^{17}$ ,  $2.5 \times 10^{17}$  and  $3 \times 10^{17} \text{ cm}^{-3}$ ).

We note that the mobility decreases with increased temperature.

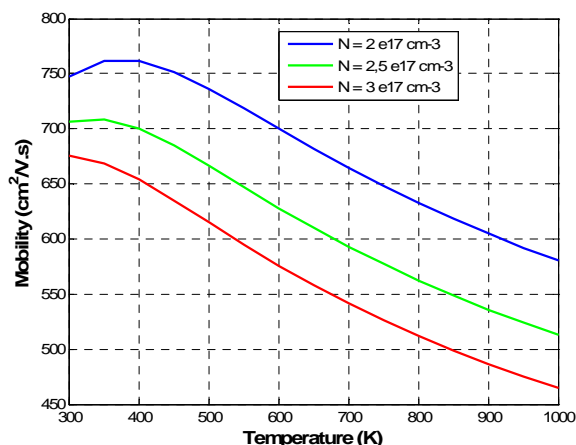


Fig. 3 Temperature effect on the electron mobility for three values doping concentrations.

The threshold voltage is the most significant parameter in study of the temperature dependence of DGMEFET characteristics. The temperature variation of threshold voltage is given by:

$$V_{th} = V_{bi} - V_p \quad (3)$$

Where  $V_p$  is the pinch-off voltage and  $V_{bi}$  is the Schottky barrier height given by [13]:

$$V_{bi}(T) = V_{bi}(T0) + m[E_{Cap}(T) - E_{Cap}(T0)] \quad (4)$$

In (4),  $m$  is between 0 and 1.

Figure 4 shows the temperature variations of threshold voltage. The threshold voltage decreases approximately linearly with increased temperature.



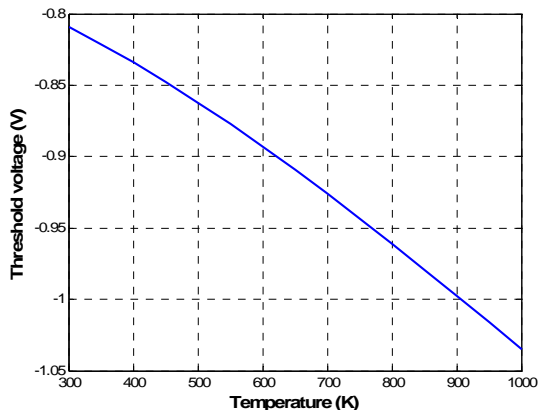


Fig. 4 Temperature dependence of the threshold voltage .

### 3. Simulated results and discussions

Figure 5 shows the studied DGMESFET structure. Figure 6 depicts the DGMESFET output  $I_{ds}$ - $V_{ds}$  characteristics for different temperatures from  $T = 300$  to  $900$  K, with a step of  $100$  K. The gates biases are  $V_{gs1} = 0.0$  V and  $V_{gs2} = -0.25$  V.

The conduction along the channel in DGMESFET is influenced by the temperature dependence of certain parameters. As shown in figure 6, the drain current decreases with increased temperature, indeed, if the temperature increases the thermal agitation of carrier's increases, this leads to decrease the mobility of carriers in the channel, and in particular to greatly affect their velocity.

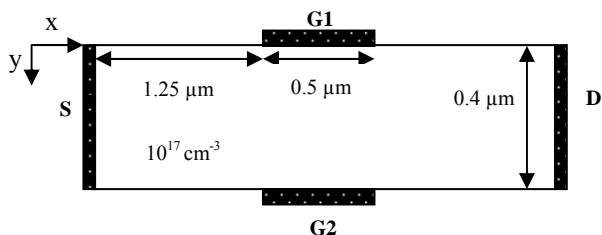


Fig. 5 Cross-section structure of the simulated DGMESFET.

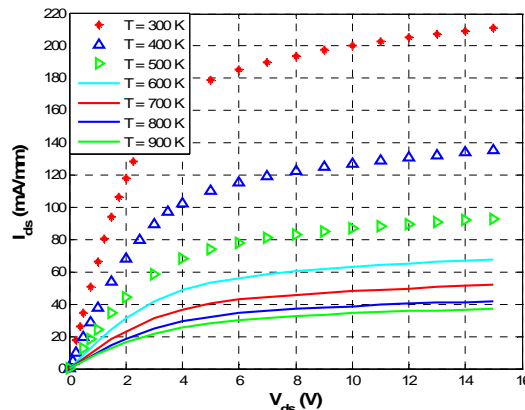


Fig. 6 DGMESFET output characteristics at  $V_{gs1} = 0.0$  V and  $V_{gs2} = -0.25$  V for different temperatures.

The transfer characteristics of the DGMESFET at different temperatures ( $T = 300$  to  $900$  K, with a step of  $100$  K), for a gate2-source voltage of  $0.0$  V and for a drain-source voltage of  $5$  V are shown in figure 7. We have used the drain bias of  $5$  V because for this polarization the transistor operates in the saturation region. Figure 7 gives an  $I_{dsmax}$  of a  $224$  mA/mm at room temperature. The threshold voltage defined as the intercept point of the gate voltage and the linear extrapolation of the  $I_{ds}$  versus  $V_{gs}$  characteristics at the maximum transconductance point, is of a  $-8$  V at room temperature.

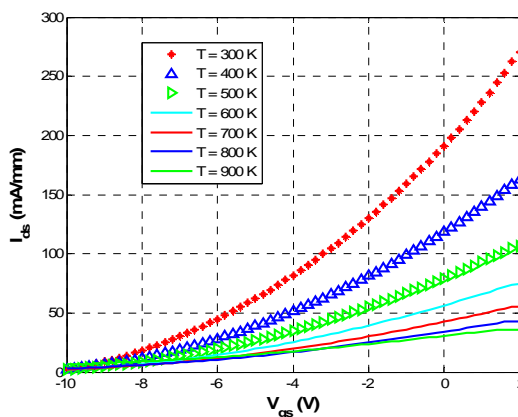


Fig. 7 DGMESFET  $I_{ds}$ - $V_{gs}$  characteristics at  $V_{ds} = 5$  V.

In order to investigate the electron temperature in the channel of the DGMESFET, figure 8 shows the variations of the electron temperature with the applied drain bias as a function of the gates bias at a point ( $x = 2.8$  μm;  $y = 0.2$  μm) located near the drain contact. As shown in figure 8, we can see that, the electron temperature is close to the lattice temperature at low drain voltages. Due to the existence of a high electric field in the gate-to-drain

region, a large electron temperature is observed when the drain voltage is greater than 1 V.

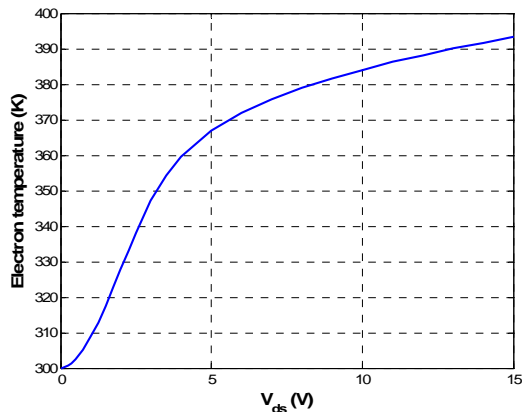


Fig. 8 Electron temperature versus drain voltage near the drain contact ( $x = 2.8 \mu\text{m}$ ;  $y = 0.2 \mu\text{m}$ ),  $V_{gs1} = V_{gs2} = 0 \text{ V}$ .

#### 4. Conclusions

In this paper, a study of the DC characteristics temperature dependence of a  $0.5 \mu\text{m}$  GaN Dual-Gate MESFET is presented. Significant physical proprieties of the device have been described using differential equations derived from a Hydrodynamic electron transport model. Two dimensional numerical simulations results over a wide range of temperatures have been presented and discussed. Degradations of the DC characteristics of the device have been observed at higher temperatures. A large electron temperature is also observed for a drain to source voltage greater than 1 V. For a low drain-source voltage ( $V_{ds} < 1 \text{ V}$ ) the electron temperature is closed to the lattice temperature.

#### References

[1] C. Lee, W. L u, E. Piner, I. Adesida, "DC and microwave performance of recessed-gate GaN MESFETs using ICP-RIE", *Solid-State Electronics*, Vol. 46, 2002, pp. 743–746.

[2] Valentin O. Turina and Alexander A. Balandin, "Electrothermal simulation of the self-heating effects in GaN-based field-effect transistors", *Journal of Applied Physics*, Vol. 100, 2006, pp. 054501.

[3] M. Schoon, "A novel, bias-dependent, small-signal model of the dual-gate MESFET", *IEEE Trans. Microwave Theory Tech.*, Vol. 42, 1994, pp.212-216.

[4] M. Ibrahim, B. Syrett, and J. Bennett, "A new analytical small-signal model of dual-gate GaAs MESFET", *IEEE MTT-S International Microwave Symposium Digest*, 2001, pp. 1277-1280.

[5] M. W. Dvorak "InAs/AlSb Heterostructure Field-Effect Transistors", Master Thesis, Simon Fraser University August, 1997.

[6] H. Djelti, M. Feham, M. Kameche, A. Ouslimani, "On the advantages of GaAs dual-gate MESFET's in comparison to Single-gate MESFET's", *2<sup>nd</sup> International Conference on Information & Communication Technologies from Theory to Applications - ICTTA'06*, 2006, Vol. 2, pp. 2562-2566.

[7] H. Djelti, M. Feham, M. Kameche, A. Ouslimani "Temperature Effect on the Drain Current of Dual-Gate GaAs MESFET's", *HITEN'2005*, France.

[8] M. Kameche, "Drain temperature determination in dual-gate GaAs MESFETs", *Journal of computational electronics*, Vol. 6, 2007, pp. 421-424.

[9] N. Lakhdar and F. Djeflal, "A two-dimensional analytical model of subthreshold behavior to study the scaling capability of deep submicron double-gate GaN-MESFETs", *Journal of Computational Electronics*, Vol. 10, No. 4, 2011, pp. 382-387.

[10] N. Lakhdar, F. Djeflal, M.A. Abdi, D. Arar, "An analytical threshold voltage model to study the scaling capability of deep submicron double-gate GaN-MESFETs", *XIth International Workshop on Symbolic and Numerical Methods, Modeling and Applications to Circuit Design*, 2010, pp. 1-4.

[11] S.-M. Sze, "Physics of Semiconductor Devices", New York: John Wiley, 1981.

[12] Farahmand, M. et. al., "Monte Carlo Simulation of Electron Transport in the III-Nitride Wurtzite Phase Materials System: Binaries and Ternaries", *IEEE Trans. Electron Devices*, Vol. 48, No. 3, 2001, pp. 535-542.

[13] M. Kameche, and N.V. Drozdovski, "GaAs-, InP- and GaN HEMT-based Microwave Control Devices: What is Best and Why", *Microwave J.*, Vol. 48, No. 5, 2005, pp. 164-180.

# Improving Multi agent Systems Based on Reinforcement Learning and Case Base Reasoning

Sara Esfandiari<sup>1</sup>, Behrooz Masoumi<sup>1</sup>, Mohammad Reza Meybodi<sup>2</sup>, Abdolkarim Niazi<sup>3</sup>

<sup>1</sup> Department of Computer Engineering and Information Technology, Islamic Azad University, Qazvin Branch, Qazvin, Iran

<sup>2</sup> Departments of Computer Engineering, Amirkabir Industrial University, Tehran, Iran,

<sup>3</sup> Department of Manufacturing and Industrial Engineering, Faculty of Mechanical Engineering, Universiti Teknologi Malaysia, 81310 UTM Skudai, Malaysia

## Abstract

In this paper, a new algorithm based on case base reasoning and reinforcement learning is proposed to increase the rate convergence of the Selfish Q-Learning algorithms in multi-agent systems. In the propose method, we investigate how making improved action selection in reinforcement learning (RL) algorithm. In the proposed method, the new combined model using case base reasoning systems and a new optimized function has been proposed to select the action, which has led to an increase in algorithms based on Selfish Q-learning. The algorithm mentioned has been used for solving the problem of cooperative Markov's games as one of the models of Markov based multi-agent systems. The results of experiments on two ground have shown that the proposed algorithm perform better than the existing algorithms in terms of speed and accuracy of reaching the optimal policy.

**Keywords:** Reinforcement learning, Selfish Q-learning, Case-base reasoning Systems, Multi-agent Systems, Cooperative Markov Games.

## 1. Introduction

Case Based Reasoning (CBR) is a knowledge based problem solving technique, which is based on reusing on the previous experiences and has been originated from the researches of cognitive sciences [1]. In this method, it is assumed that the similar problems can possess similar solutions. Therefore, the new problems may be solvable using the experienced solutions to the previous similar problems. A multi-agent system (MAS) is comprised of a collection of intelligent agents that interact with each other in an environment to optimize a performance measure [2].

Agents are computational entities that can see their environments with their sensors. These agents should do appropriate action in per moment based on their observations. In multi agent system research, cooperative and non-cooperative perspective. In cooperative multi-agent systems, the agents pursue a common goal and the agents can be built expect benevolent intentions from other agents. In contrast, a non-cooperative multi agent system setting has non-aligned goals, and individual agents try to obtain only to maximize their own profits. In multi-agent systems, the need for learning and adoption is essentially caused by the fact that the environment of the agent is dynamic and just empirically observed while the environment (the reward functions and the transition states) is unknown. Hence, the reinforcement learning methods may be applied in MAS to find an optimal policy in MGs. In addition, agents in a multi-agent system face the problem of incomplete information with respect to the action choice. If agents get information about their own choice of action as well as that of the others, then we have joint action learning [3][4]. Joint action learners are able to maintain models of the strategy of others, and the explicitly takes into account the effects of joint actions. In contrast, independent agents only know their own action which is often a more realistic assumption since distributed multi-agent applications are typically subject to limitations such as partial observability, communication costs, and stochastic.

There are several models proposed in the literatures for multi-agent systems based on Markov models. One of these models is stochastic games (also called Markov Game – MG). Markov games are extensions of Markov

Decision Process (MDP) to multiple agents. In an MG, actions are the result of joint action selection of all agents, while rewards and the state transitions depend on these joint actions. In a fully cooperative MG called a multi-agent MDP (or MMDP), all agents share the same reward function and they should learn to agree on the same optimal policy [5].

There are several methods for finding an optimal policy in MMDPs. In [6], an algorithm is proposed for learning cooperative MMDPs, but it is only suitable for deterministic environments. In [7] an algorithm called *Selfish Q-Learning* has been introduced which changes the Q values of each action used a special Q-function. In [8] MMDPs are approximated as a sequence of intermediate games. The authors present optimal adaptive learning and prove convergence to Nash equilibrium of the game. In [9], an algorithm called *CAQL* has been introduced, which acts through a *Q - learning algorithm*. In [10], a Q-learning algorithm based method has been proposed.

In Reinforcement Learning (RL), learning is carried out online, through trial-and-error interactions of the agent with the environment. Unfortunately, convergence of any RL algorithm may only be achieved after extensive exploration of the state-action space, which can be very time consuming. However, the rate of convergence of an RL algorithm can be increased by using heuristic functions for selecting actions in order to guide the exploration of the state-action space in a useful way. In [11], [12] investigates how to make improved action selection functions based on heuristics in on-line policy learning for robotic scenarios. These functions have been applied to select the action in every state. Although these methods have been successfully used to find the optimal policy in Markov games, the problem of using the previous experiences of agents for solving the new problem is still disregarded in these methods. Since in the environment is unknown in multi-agent systems, and the agent should upgrade its knowledge of environment through observation, so the problem of keeping and reusing the previously acquired knowledge causes an increase in learning rate. In this paper, to increase the speed of learning rate to get the optimal policy for Markov Games in the independent agent's state, a hybrid algorithm called Case-based Best Heuristically Accelerated Selfish Q-learning (CB-BHASQL) is proposed in which, a modified function is used to select the action and the Case Base Reasoning technique and a special Q-function called Selfish Q-Learning has been used to increase the learning rate. To evaluate the proposed methods, they have been applied to two examples of MMDP called Grid Game and Tunnel To Goal. The results of computer simulations have shown that these algorithms outperform the previous approaches from both cost and speed perspective. In the next part of the paper, at first fundamental concepts are

explained in section 2 and in section 3, the proposed algorithm is presented. Simulation results, and discussions are reported in section 4 and in section 5, evaluation of the algorithm's behavior and its analysis is done and section 6 is the conclusion.

## 2. Reinforcement Learning

In this section, we first review some basic principles of Markov decision Process (MDP) and then present the basic formulation of the Q-learning algorithm, a well-known reinforcement learning technique for solving MDPs. A reinforcement learning agent defines its behavior through interaction with an unknown environment and observation of the results of its behavior [12].

### 2.1 Markov decision Process

Markov decision process is formally defined as follows:

**Definition 1.** A Markov decision process (MDP) is a quadruple  $\langle S, A, R, T \rangle$  (where  $S$  is a finite state space;  $A$  is the space of actions the agent can take;  $R: S \times A \rightarrow \mathcal{R}$  is a payoff function ( $R(s, a)$  is the expected payoff for taking action  $a$  in state  $s$ ); and  $T: S \times A \times S \rightarrow [0,1]$  is a transition function ( $T(s, a, s')$  is the probability of ending in state  $s'$ , given that action  $a$  is taken in state  $s$ ).

In a Markov decision process, an agent's objective is to find a strategy (policy)  $\pi: S \rightarrow A$  so as to maximize the sum of discounted expected rewards,

$$V(s, \pi) = \sum_{t=0}^{\infty} \gamma^t E(r_t | \pi, s_0 = s) \quad (1)$$

Where  $s$  is a particular state,  $s_0$  indicates the initial state,  $r_t$  is the reward at time  $t$ , and  $\gamma \in [0,1)$  is the discount factor. There exists an optimal policy  $\pi^*$  such that for any state  $s$ , the following equation holds:

$$V(s, \pi^*) = \max_a \{r(s, a) + \gamma \sum_{s'} P(s'|s, a) v(s', \pi^*)\} \quad (2)$$

where  $r(s, a)$  is the reward for taking action  $a$  at state  $s$ , and  $v(s, \pi^*)$  is called *optimal value* for that state while  $P(s'|s, a)$  is the probability of transiting to state  $s'$  After taking action  $a$  in state  $s$ . If the agent knows the reward and state transition functions, it can solve  $\pi^*$  by iterative search method, otherwise this method cannot be used while an algorithm called *Q* is employed [13] [14]. The variety of Q-functions has been used in [6,7,8,9,10]. Selfish Q-learning algorithm pseudo-code, which has been used in [7], is shown in Figure 1. In this algorithm, for every action  $a$  in each state  $S$  the value of that action ( $Q(s, a)$ ) is used according to Equation 3. Each state  $S$  the value of that action ( $Q(s, a)$ ) is used according to

Equation 3. In Equation3,  $\alpha$  is the rate of learning and  $\gamma \in [0, 1]$  is the discount factor. The algorithm ends when the optimum policy doesn't change for a definite while.

$$Q(S_t, a_t) = Q(S_t, a_t) + \alpha [r_t + (\gamma \max_{a'} Q(S', a')) - Q(S_t, a_t)] \quad (3)$$

To select an action in every state, the Boltzmann's distribution method (EQ 4) is usually used. This function has been used in some articles such as [6,7,8,9,10,19].

$$\pi_1(S) = \arg \max \left( \frac{e^{\frac{Q(S,a)}{\tau}}}{\sum_{i=1}^m e^{\frac{Q(S,i)}{\tau}}} \right) \quad (4)$$

In which,  $m$  is the number of allowable actions for state  $S$  and  $\tau$  is a constant.  $Q(S, a)$  shows the value of evaluation function of state  $S$  while action  $a$  is done.

```

Algorithm Selfish Q-Learning
1. Initialize Q(S,a) arbitrarily
2. Repeat (for each episode)
3. Initialize S randomly
4. Repeat (for each step)
5. Select an action using  $\pi_1(S) = \arg \max \left( \frac{e^{\frac{Q(S,a)}{\tau}}}{\sum_{i=1}^m e^{\frac{Q(S,i)}{\tau}}} \right)$  EQ(4)
6. Execute the action a
7. Observe reward  $r(s,a)$ , state  $s'$ 
8. Update the value of  $Q(S,a)$  according to  $Q(S_t, a_t) = Q(S_t, a_t) + \alpha [r_t + (\gamma \max_{a'} Q(S', a')) - Q(S_t, a_t)]$  EQ(3)
9.  $S \leftarrow S'$ 
10. Until S is Terminal State
11. Until some stopping Criteria is reached.
12. End
    
```

Fig 1. Selfish Q-Learning Algorithm

## 2.2. Markov Games

Markov games are a generalization of MDPs to multiple agents and can be used as a framework for investigating multi-agent learning. In the general case (general-sum games), each player would have a separate payoffs. A standard formal definition follows:

**Definition 2.** A stochastic game (Markov game) is a tuple  $\langle n, S, A_{1..n}, T, R_{1..n} \rangle$ , where  $n$  is the number of agents,  $s$  is a set of states,  $A_i$  is the set of actions available to agent  $i$  (and  $A$  is the joint action space  $A_1 \times A_2 \times \dots \times A_n$ ),  $T$  is a transition function  $S \times A \times S \rightarrow [0,1]$ , and  $r$  is a reward function for the  $i_{th}$  agent  $S \times A \rightarrow \mathcal{R}$ .

In a discounted Markov game, the objective of each player is to maximize the discounted sum of rewards, with a discount factor  $\gamma \in [0,1)$ . Let  $\pi_i$  be the strategy of the player  $i$ . For a given initial state  $s$ , player  $i$  tries to maximize:

$$v(s, \pi^1, \pi^2, \dots, \pi^n) = \sum_{t=0}^{\infty} \gamma^t E(r_t | \pi^1, \pi^2, \dots, \pi^n, s_0 = s) \quad (5)$$

Markov games are categorized based on the agent's rewards into cooperative and non-cooperative games. Non-cooperative games may be classified as competitive games and general-sum games. Strictly competitive games, or zero-sum games, are two-player games where one player's reward is always the negative of the others. General-sum games are ones where the reward sum is not restricted to zero or any constant, and allow the agents' rewards to be arbitrarily related. However, in full cooperative games, or team games, rewards are always positively related. In a fully cooperative MG (or team MG) called a multi-agent MDP (or MMDP), all agents share the same reward function. Nevertheless, in general MG (or general-sum MG) there is no constraint on the sum of the agents' rewards and the agents should learn to find and agree on the same optimal policy. However, in a general Markov Game, an equilibrium point is sought; i.e. a situation in which no agent alone can change its policy to improve its reward when all other agents keep their policy fixed [15], [16].

One of the Markov's games used for multi-agent Markov's games is the Grid World game. In this game, two agents start from a corner of the page and try to reach a goal with the least possible number of moves. Players' actions are defined as four actions in four different directions, namely Up, Down, Left, Right. A state space set is defined as  $S = \{s | s = (l_1, l_2)\}$ , In which each state  $s = (l_1, l_2)$  Indicates the coordinates of agents 1 and 2. Agents cannot take the same coordinates at the same time. In other words, if both agents try to move to the same square, both of their moves will fail. If agents move to two different non-goal positions, both receive zero rewards and if one reaches the goal position, it receives 100 units of reward. However, if they collide with each other both receive one unit of punishment and stay in their previous position. In this game, the state transition is deterministic, i.e. the next state is uniquely determined by the current state and the joint action of the agents. In this game, agents are assumed not to know the goal position and the other agent's reward functions. Agents choose their actions simultaneously and can only know about the previous moves of the other agents and their own current state.

Another game which we have used it, called *Tunnel to Goal*. In this game, there are some barriers. If an agent collision these barriers, it receives one unit of punishment.

A path in these games represents sequences of actions from the starting to the end position. In game terminology,

such a path is called a policy or strategy. The shortest path, not interfering the path taken by the other agent, is called the optimal policy or *Nash path*. Figure 2 is an example of these games. The optimal policy in Figure 2.a includes 9 movements and in Figure 2.b includes 10 movements.

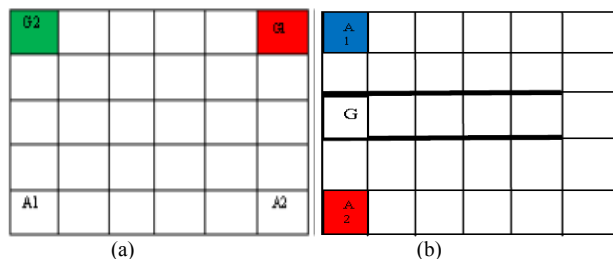


Fig2 . examples of Markov Games. (2.a) An example of Grid World Game. 2.b. An example of Tunnel to Goal Game

### 2.3. Case Base Reasoning

Case Based Reasoning (CBR) technique uses the previous experiences (Case) to solve the new problems [17], [18]. In the case base reasoning systems, the experiences gained from solving the problems are saved in case base (CB). In these systems, for solving the new problem ( $C_{new}$ ), the most similar cases to  $C_{new}$  are extracted from the case base (CB) and the solutions presented by the extracted cases are used to solve the new problem  $C_{new}$ . If a similar case is not found,  $C_{new}$  is inserted to the case base as a new case. Unlike the classical knowledge-based methods, CBR focuses on a particular problem-solving experience, which is originated from the cases collected in the case base. These cases show a particular experience on a problem solving domain. It must be noted that CBR doesn't recommend a definite solution, but presents hypothesis and theories pass the solution space.

### 3. The Proposed method

In this section, a new algorithm called *CB-BHASQL* is proposed to increase the rate of convergence in Markov's games. In the proposed algorithm, the case base reasoning and also a new function are used to select the action in each state to increase the convergence rate toward the optimal policy. We previously used our new function with Decentralized Q-Learning according to EQ(6) and called CB-BHADQL. In this paper, We proposed a new

algorithm with Selfish Q-Learning according to EQ(3) and called CB-BHASQL.

$$Q(S, a) = (1 - \alpha)Q(S, a) + \alpha(r + \gamma \max_b Q'(S, a)) \quad (6)$$

We know that solving a problem using CBR includes the steps: creating a description of the problem, evaluating the similarity of the current problem to the previously-solved problems saved in case base, and trying to reuse the solutions presented by the detected cases to solve the current problem. The structure of the cases used in the recommended algorithm is a duplex in the form of  $Case = \langle Prob, Sol \rangle$  in which, *Prob* describes the problem and *Sol* is the solution presented to solve the problem. The problem describer (*Prob*) includes the properties in each state. In the proposed algorithm, the problem describer is defined as  $Prob(S) = \{m, \langle Up, Down, Right, Left \rangle, index\}$  in which,  $m$  is the number of actions for each state and the set  $\langle Up, Down, Right, Left \rangle$  are the actions allowable for each action and *index* is the index for each state. The solution recommended for the problem is  $Sol(S) = \langle E, V \rangle$ , in which vector  $\vec{E}$  In the form of  $\vec{E} = (\vec{E}[1], \vec{E}[2], \dots, \vec{E}[m])$  is a list of experiences collected from the environment by the agent for state  $S$  and each vector  $\vec{E}$  includes a tuple  $\langle A_i, N_i, Q_i, \pi_i \rangle$  where  $A_i$  is the space of actions for state  $S$  and  $N_i$  is the number of times that  $a_i \in A_i$  has been updated and  $Q_i$  is the value estimated by Equation 1 and  $\pi_i$  is the possibility of occurrence of action  $a_i$ , which is estimated by EQ(7).

$$\pi_2(S) = \arg \max \left( \frac{e^{n(S,a)Q(S,a)}}{\sum_{i=1}^m e^{n(S,a)Q(S,a)}} \right) \quad (7)$$

Where  $m$  is the number of allowable actions for state  $S$  and  $n(S, a)$  is the number of times that so far the action  $a$  has been selected.  $Q(S, a)$  shows the value of evaluation function of state  $S$  while action  $a$  is done.

$V$  is the justification of using the solution recommended by the detected agent and if each of the actions of state  $S$  at least has been selected once, the solution of the detected agent can be used for solving the new problem. In the recommended algorithm, once the agent enters a new state, extracts the most similar case to the new state of the case

base and if the justification is available ( $V = True$ ), the detected case is used to determine the next state.

To detect the similar cases of current state, the nearest neighbor algorithm is used. Euclidean distance of the new case to each of the cases available in case base is calculated according to EQ(8) and the most similar case ( $c$ ) is detected and if the justification is available ( $V = True$ ), the solution of detected case is used to solve the new problem. The proposed algorithm is shown in Figure 3.

$$NN(S) = \arg \max_{c \in CB} Sim(C. prob, C. sol) \\ = \arg \max_{c \in CB} dist(C. prob, C. sol) \quad (8)$$

## 4. Experiments

In order to evaluate the performance of the proposed algorithm several experiments have been conducted whose results are reported below. In Section 4.1, The environment of the experiments is a Grid-world game that includes a  $5 \times 6$  Grid according to Figure 2.a. In section 4.2, the environment of the experiments is a Tunnel to Goal game that includes a  $5 \times 6$  Grid according to Figure 2.b. These experiments are conducted to study the improvement obtained by the proposed algorithm (CB-BHASQL) in comparison with three CBR and QL algorithms and CB-BHADQL. So, CB-BHASQL algorithm is compared with three algorithms: 1) Selfish Q-Learning algorithm, and 2) Boltzmann's CBR algorithm, which its pseudo-code is similar to Figure 3 and the only difference is in the selection of the actions which is based on the Boltzmann's distribution (Equation 4) and 3) CB-BHADQL. In all experiments, each reported value is obtained by averaging over 200 runs and the average results are gained for the algorithms. Parameters given are  $\tau = 0.05$  and  $\gamma = 0.7$ .

### 4.1. Experiments in Grid World Games

In this section, we show results of our experiments in Grid word Games.

**Experiment 1.** In this experiment, we compare the proposed algorithm (CB-BHASQL) in Grid World Games environment ( Fig 2.a) with the other algorithms in terms of the number of movements made by agent 1 to reach the optimal path in 2000 episode. Figure 4 illustrates the results of this experiment. Figure 5 shows the average results after 200 runs. From the result, it is evident that the

CB-BHASQL algorithm has lower numbers of moves in comparison with the other algorithms.

**Experiment 2.** In this experiment, we compare the proposed algorithm (CB-BHASQL) in Grid World Games environment ( Fig 2.a) with the other algorithms in terms of the averaged reward received by agent 1 during an episode. Figure 6 shows the result of this experiment. As it is seen (CB-BHASQL) algorithm outperforms the other algorithms in terms the average reward received during an episode.

**Algorithm CB-BHASQL**

1. Let  $t$  be the global time,  $n$  be the number of agents,  $\gamma$  the discount factor,  $CB_i = \phi$  an empty case base for each Set  $s = s' \in Sto$  the initial state of the system
2. **Repeat**
3. Set  $s = s'$
4. **forall** agent  $i \in [1 \dots n]$  **do**
5.     **if**  $CB_i = \phi$  or **addcasecriterion(s) is true**
6.          $CB = CB \cup C$  with  $c.Prob = s$  and  $c.Sol = empty\_solution(i)$
7.         **for each**  $j = Sol(s).m$  **do**
8.             **Compute**  $Sol(s).E[j].\pi_i$  according to EQ(7) and **Set** index  $X_i$  the *Maximum value of them.*
9.             **Select elementary action**  $Sol(s).E[x_i].a_i$ .
10.     **Observe Successor state**  $s' \in S$  and **reward**  $r \in R$ .
11.     **end if.**
12.     **end for.**
13.     **for all** agents  $i \in [1 \dots n]$  **do**
14.         **Retrieve nearest neighbour** according to EQ(8) of state  $s'$ .
15.         **Set Learning Rate**  $\alpha_i = \frac{1}{1 + Sol(s).E[x_i].n_i}$
16.         **Set**  $Sol(s).E[x_i].Q_i$  according to  $Q(S_t, a_t) = Q(S_t, a_t) + \alpha[r_t + (\gamma \max_{a'} Q(S', a')) - Q(S_t, a_t)]$  EQ(3).
17.         **Increment**  $Sol(s).E[x_i].n_i$  by one.
18.         **Resort** decrementsly the experience list  $Q$  in  $Sol(s).E$
19.     **Until**  $Stop\_Criterion()$  becomes true.

Fig 3. Pseudo-code for the Proposed Algorithm CB-BHASQL

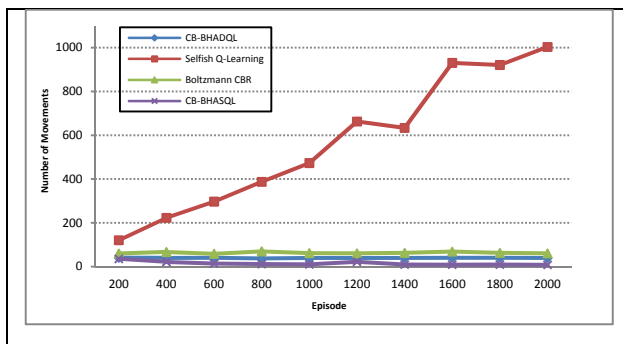


Fig4. Comparison of different methods in terms of the number of movements Needed for reaching to the optimal path in Grid World.

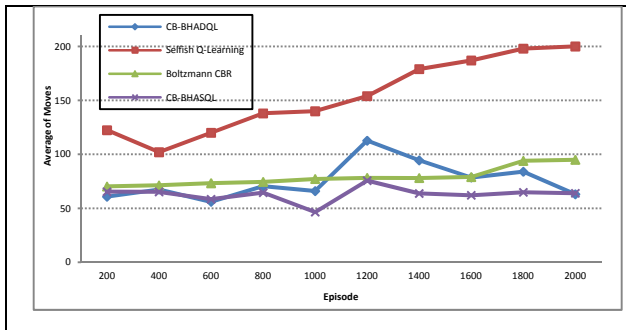


Fig5. Comparison of four Algorithms in terms of average of the number of moves to reaching optimal path in 200 runs in Grid World.

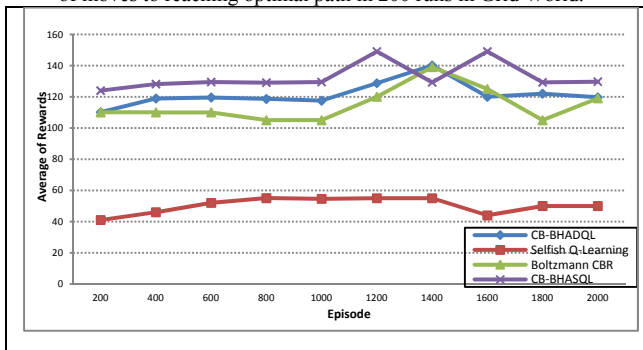


Figure 6. Comparison of Three Algorithms in term of the average of rewards gained in 200 runs in Grid World.

## 4.2. Experiments in Tunnel to Goal Games

In this section, we show results of our experiments in Tunnel to Goal Games.

**Experiment 3.** In this experiment, we compare the proposed algorithm (CB-BHASQL) in Tunnel to Goal Games environment ( Fig 2.b) with the other algorithms in terms of the number of movements made by agent 1 to reach the optimal path in 2000 episode. Figure 7 illustrates the results of this experiment. Figure 8 shows the average results after 200 runs. From the result, it is evident that the CB-BHASQL algorithm has lower numbers of moves in comparison with the other algorithms.

**Experiment 4.** In this experiment, we compare the proposed algorithm (CB-BHASQL) in Tunnel to Goal Games environment ( Fig 2.b) with the other algorithms in terms of the averaged reward received by agent 1 during an episode. Figure 9 shows the result of this experiment. As it is seen (CB-BHASQL) algorithm outperforms the other algorithms in terms the average reward received during an episode.

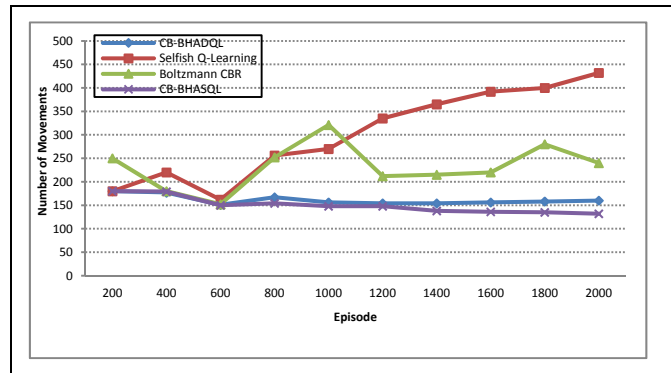


Fig7. Comparison of different methods in terms of the number of movements Needed for reaching to the optimal path in Tunnel to Goal.

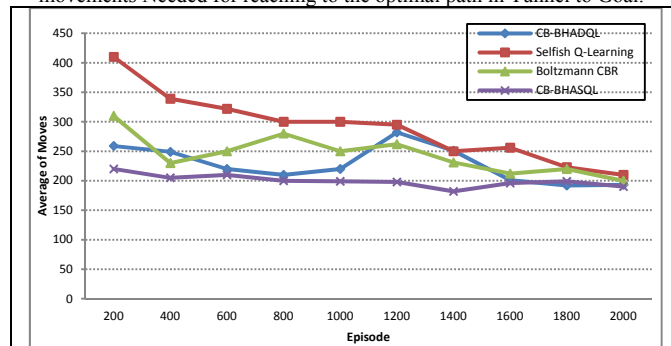


Fig8. Comparison of four Algorithms in terms of average of the number of moves to reaching optimal path in 200 runs in Tunnel to Goal.

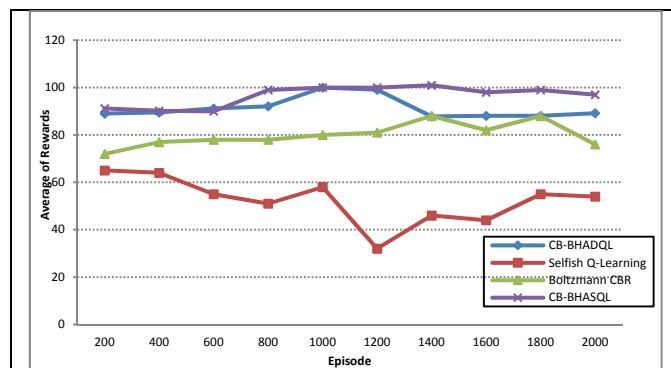


Fig 9. Comparison of Three Algorithms in term of the average of rewards gained in 200 runs in Tunnel to Goal.

## 5. Evaluation of the Algorithm's Behavior

### 5.1. Examination of the Behavior of the Proposed Algorithms

In this section, an analysis of the performance the proposed algorithm is conducted in which the advantage of the function  $\pi_2(S)$  (EQ 7) is compared with  $\pi_1(S)$  (EQ 4). We want to show that in the proposed method,  $\pi_2(S)$  in comparison with  $\pi_1(S)$  converges to the optimum solution with a higher rate. In other words, the rate of variation for



$\pi_2(S)$  in relation to  $Q$ , is more than the rate of variation for  $\pi_1(S)$  in relation to  $Q$ .

To show the advantage of the behavior of the proposed action selection function, the *CB-BHASQL* algorithm was evaluated for state  $S_0$  and action  $a_1$  regarding to different values of  $n$  and the results below were gained:

**Experiment 5.** In this experiment, variations of  $\pi_2(S)$  were evaluated in comparison with  $\pi_1(S)$ . Figures 10-12 show these variations. As it is seen, we note that regarding to the increase in  $n$ , the growth of  $\pi_2(S)$  is much more than  $\pi_1(S)$ .

**Experiment 6.** In this experiment, we study evaluation of variations for function  $Q(S, a)$  (EQ(3)) regarding to the increase in  $n$ . The results of this analysis are shown in Figures 11-13. As it is seen, we conclude that with increasing  $n$ , the value of function  $Q(S, a)$  also increases in Equation 3.

**Experiment 7.** In this experiment, we study evaluation of variations for  $\pi_1(S)$  and  $\pi_2(S)$  based on values for  $Q(S, a)$ . The results of this evaluation are shown in Figure 14. Looking at the diagram we note that with increasing value of  $Q(S, a)$ , the value of  $\pi_1(S)$  increases. Since always  $\lim_{t \rightarrow \infty} \text{not}(S, a) = \infty$ , according to the result of Experiment 6, value of  $Q_t(S, a)$  increases and according to the result of Experiment 6, with increasing  $n$ , the function  $\pi_2(S)$  grows faster than  $\pi_1(S)$ . Based on the previous subjects, it is concluded that with increasing value of  $Q_t(S, a)$ , the function  $\pi_2(S)$  must grow faster than  $\pi_1(S)$ . Figure 15 shows the results.

## 5.2. Mathematical Analysis of the Functions Behavior

To facilitate the calculations, we rewrite functions  $\pi_1(S)$  and  $\pi_2(S)$  as EQ (9) and EQ (10) respectively.

$$\pi_1(S) = \frac{e^{\frac{Q}{\tau}}}{\sum_{j=1}^m e^{\frac{Q_j}{\tau}}} \quad (9)$$

$$\pi_2(S) = e^{nQ} \quad (10)$$

The variable rate of  $\pi_1(S)$  in relation to  $Q$  with parameter  $t = 0.05$ , is shown in EQ (11).

$$\frac{\Delta\pi_1(S)}{\Delta Q} = \frac{1}{\tau} e^{\frac{1}{\tau}} = \frac{1}{0.05} e^{\frac{1}{0.05}} = 20e^{20Q} \quad (11)$$

The variable rate of  $\pi_1(S)$  in relation to  $Q$  is shown in EQ (12).

$$\begin{aligned} \frac{\Delta\pi_2(S)}{\Delta Q} &= \frac{d\pi_2(S)}{dn} \times \frac{dn}{dQ} \\ &= Qe^{nQ} \times \frac{dn}{dQ} \end{aligned} \quad (12)$$

Through the comparison of the EQ(11) and EQ(12), we conclude that the growth of the rate  $\frac{\Delta\pi_1(S)}{\Delta Q}$  is less than  $\frac{\Delta\pi_2(S)}{\Delta Q}$ .

In Equation 11, because  $Q$  is positive, the value of function  $\frac{\Delta\pi_1(S)}{\Delta Q}$  is always positive. According to equation 3 and the diagram in Figure 11, with increasing  $n$ , the value of  $Q(S, a)$  always increases. Thus,  $\frac{dn}{dQ} > 0$  and from the other hand,  $n > 0$  and  $Q > 0$ . So,  $\frac{\Delta\pi_2(S)}{\Delta Q}$  is always positive.

In Equation 10,  $e^{20Q}$  is multiplied by the constant value 20. This is while in Equation 12,  $e^{nQ}$  is multiplied by variant  $Q$ . With increasing  $n$ , the value of  $Q$  increases. Thus, the growth rate of  $\frac{\Delta\pi_1(S)}{\Delta Q}$  is less than  $\frac{\Delta\pi_2(S)}{\Delta Q}$ . The diagram in Figure 15 also supports the result gained.

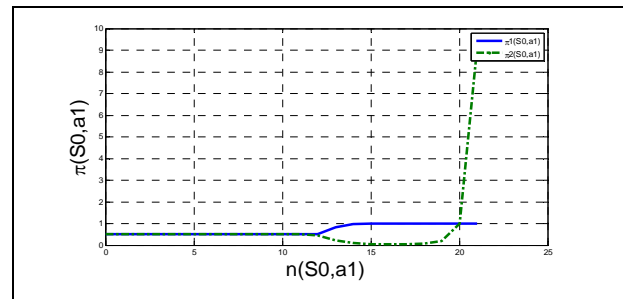


Fig 10. Examining the growth of  $\pi_2(S)$  and  $\pi_1(S)$  based on the different  $n$  values

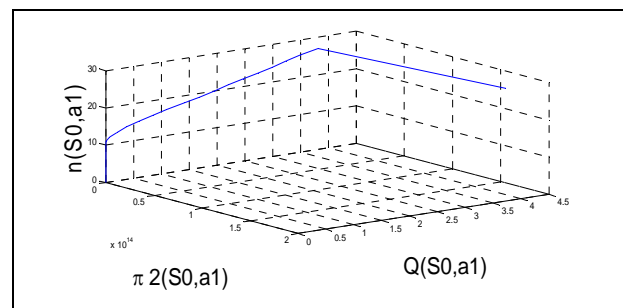


Fig 11. Examination of  $\pi_2(S)$  variations based on the different  $n$  values.

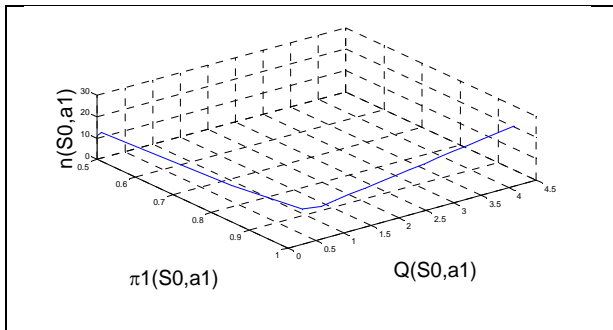


Fig 12. Examination of  $\pi_1(S)$  Variations based on the Different  $n$  values.

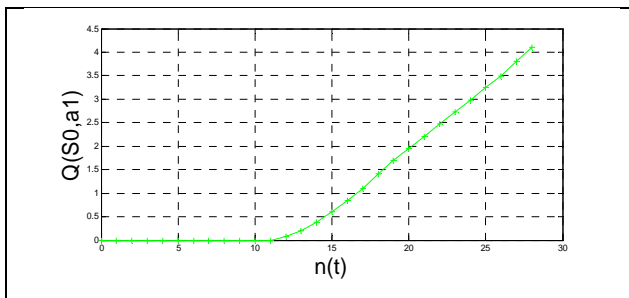


Fig 13. Examination of  $Q(S, a)$  Variations based on the Different  $n$  values.

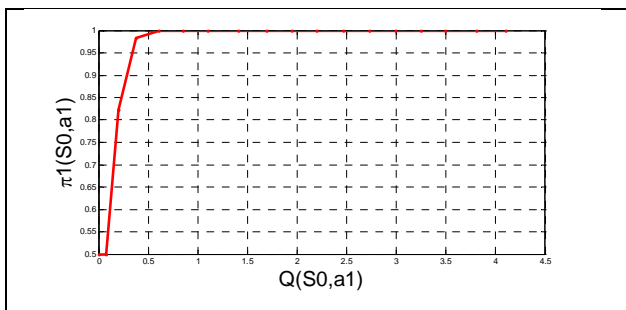


Fig 14. Examination of  $\pi_1(S)$  Growth based on the Different  $Q(S, a)$  values.

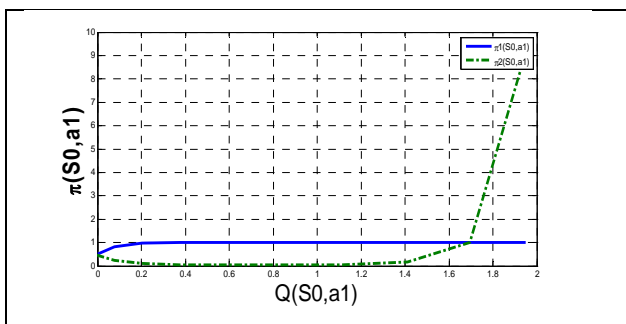


Fig 15. Examination of  $\pi_1(S)$  and  $\pi_2(S)$  Growth based on the Different  $Q(S, a)$  values

## 6. Conclusion

In this paper, a new hybrid model called *CB-BHAQL* has been introduced to solve Markov's games based on reinforcement learning and case base systems, in which a

new function has been applied to select the action in each state. The results gained were compared with the current algorithms. Based on the results gained in two different environments, in comparison with *Selfish Q-Learning* and *Boltzmann's CBR* algorithms and *CB-BHADQL*, our proposed algorithm *CB-BHASQL* algorithm has a very high efficiency from the perspective of convergence rate to the optimum answer, average of total reward gained and the number of the movements needed for convergence to the optimal policy.

## References

- [1] R. A. C. Branchi, R. Raquel, R. L. D. Mantaras, "Improving Reinforcement Learning by using Case Based Heuristics", Proceeding of the Int. Conference on Case Based Learning 2009 (ICCBRL 2009), Springer, 2009.
- [2] N. Vlassis, "A Concise Introduction to Multiagent Systems and Distributed Artificial Intelligence", 2007, Morgan and Claypool Publishers.
- [3] C. Boutilier, "Sequential optimality and coordination in multi-agent systems", in: Proceedings of the 16th International joint conference on Artificial intelligence, 1999, Vol. 1, Morgan Kaufmann Publishers Inc., Stockholm, Sweden.
- [4] L. Bosni, R. Babuska, and B. Schutter, "A Comprehensive Survey of Multiagent Reinforcement Learning", IEEE Transaction on System, Man, Cybern, 2008, vol. 38, pp. 156-171.
- [5] B. Masoumi, M. R. Meybodi, "Speeding up learning automata based multi agent systems using the concepts of stigmergy and entropy", Journal of Expert Systems with Applications, July 2011, Vol 38, Issue 7, PP. 8105-8118.
- [6] M. Lauer and M. Riedmiller, "An Algorithm for Distributed Reinforcement Learning in Cooperative Multi-Agent Systems", in The 17<sup>th</sup> International Conference on Machine Learning San Francisco, CA, USA, 2000: Morgan Kaufmann Publishers Inc, pp. 535 – 542.
- [7] A. G. Barto and R. S. Sutton, "Reinforcement Learning: an introduction", MIT Press, Cambridge, MA, 1998.
- [8] X. Wang and T. Sandholm, "Reinforcement Learning to Play an Optimal Nash Equilibrium in Team Markov Games", in Advances in Neural Information Processing Systems, 2002, vol. 15: MIT Press, pp. 1571-1578, 2002,
- [9] F. S. Melo, M. I. Ribeiro, "Reinforcement Learning with Function Approximation for Cooperative Navigation Tasks", IEEE International Conference on Robotics and Automation Pasadena, CA, USA, May 2008, pp. 3321-2237.
- [10] M. Lauer and M. Riedmiller, "Reinforcement Learning for Stochastic cooperative Multi-agent Systems", In Proceeding of AAMAS 2004, New York, NY, ACM Press, pp. 1514-1515.
- [11] R. A. C. Bianchi, C. H. C. Ribeiro, A. H. R. Costa, "Accelerating autonomous learning by using a heuristic selection of actions", Journal of Heuristics, 2008, Vol. 2, pp.135-168.
- [12] R. A. C. Bianchi, C. H. C. Ribeiro, A. H. R. Costa, "Heuristic selection of actions in multi agent reinforcement

- learning”, 20<sup>th</sup> International conference on Artificial Intelligence, India, Jan 2007, pp.690-695.
- [13] L. Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming, John Wiley and Sons, New York, 1994.
- [14] R. S. Sutton, A. G. Barto, “Reinforcement Learning : An Introduction”, MIT Press, 1998.
- [15] J. F. Nash, “Non-cooperative Games”, Annals of Mathematics, , 1951, Vol. 54, pp. 286–295.
- [16] A. M. Fink, Equilibrium in a Stochastic N-person Game, Journal of Science in Hiroshima University, Series A-I, 1964, Vol. 28, pp. 89–93.
- [17] A. Aamodt; E. Plaza, "Case-Based Reasoning: Foundational Issues", Methodological Variations and System Approaches AI Communications, IOS Press, 1994, Vol. 7, No. 1, pp. 39-59.
- [18] R. Bergman; "Engineering Applications of Case Based Reasoning", Journal of Engineering Applications of Artificial Intelligence, 1999 , Vol. 12, pp.805.
- [19] Gabel, T. And Riedmiller, M., “CBR for state value function Approximation in Reinforcement Learning”, Proceeding of the Inter. Conference on Case Based Learning 2005 (ICCBRL 2005) , Springer , Chicago, USA.

**Sara Esfandiari** received her BS degree in Computer Engineering from the azad University, Tehran, Iran., in 2006 and MS degree in Computer Engineering in 2011 from the azad University, Qazvin, Iran. Her research interests include Learning systems, multi agent systems, multi agent learning, Data Mining, parallel algorithms.

**Behrooz Masoumi** received his BS and MS degrees in Computer Engineering in 1995 and 1998, respectively. He also received his PhD degrees in Computer Engineering from the Science and Research University, Tehran, Iran., in 2011. He joined the faculty of Computer and IT Engineering Department at Qazvin Azad University, Qazvin, Iran, in 1998. His research interests include learning systems, multi-agent systems, multi-agent learning, and soft computing.

**Mohammadreza Meybodi** received his BS and MS degrees in Economics from the Shahid Beheshti University in Iran, in 1973 and 1977, respectively. He also received his MS and PhD degrees in Computer Science from the Oklahoma University, U.S.A., in 1980 and 1983, respectively. Currently he is a Full Professor in the Computer Engineering Department, Amirkabir University of Technology, Tehran, Iran. Prior to his current position, he worked from 1983 to 1985 as an Assistant Professor at Western Michigan University, and from 1985 to 1991 as an Associate Professor at Ohio University, U.S.A. His research interests include channel management in cellular networks, learning systems, parallel algorithms, soft computing, and software development.

**Abdolkarim Niazi** is currently PhD Student in Mechanical Engineering-Manufacturing engineering at Technical University of Malaysia from 2010 up to now. His research interests include Condition monitoring, Tools Condition monitoring, Tool wear and tool vibration, Advance and Automated Manufacturing Systems, Artificial Neural Networks.

# Peer to Peer Networks Management Survey

Mourad AMAD<sup>1</sup>, Ahmed MEDDAHI<sup>2</sup> and Djamil AÏSSANI<sup>3</sup>

<sup>1,3</sup> Laboratory LAMOS, University of Bejaia, Algeria

<sup>2</sup> Institut Telecom/Telecom Lille 1, France

## Abstract

Peer-to-Peer systems are based on the concept of resources localization and mutualisation in dynamic context. In specific environment such as mobile networks, characterized by high variability and dynamicity of network conditions and performances, where nodes can join and leave the network dynamically, resources reliability and availability constitute a critical issue. The resource discovery problem arises in the context of peer to peer (P2P) networks, where at any point of time a peer may be placed at or removed from any location over a general purpose network. Locating a resource or service efficiently is one of the most important issues related to peer-to-peer networks. The objective of a search mechanism is to successfully locate resources while incurring low overhead and low delay. This paper presents a survey on P2P networks management: classification, applications, platforms, simulators and security.

**Keywords:** P2P, Routing, Complexity, Algorithm, Design, performance.

## 1. Introduction

Peer-to-Peer (P2P) systems are distributed systems without (*or with a minimal*) centralized control or hierarchical organization, where each node is equivalent in term of functionality. P2P refers to a class of systems and applications that employ distributed resources to perform a critical function such as resources localization in a decentralized manner. The main challenge in P2P computing is to design and implement a robust distributed system composed of distributed and heterogeneous peer nodes, located in unrelated administrative domains. In a typical P2P system, the participants can be “domestic” or “enterprise” terminals connected to the Internet.

Peer-to-Peer computing is a very controversial topic. Many experts believe that there is not much new in P2P. There are several definitions of P2P systems that are being used by the P2P community. As defined in [48],

“P2P allows file sharing or computer resources and services by direct exchange between systems”, or “allows the use of devices on the Internet periphery in a non client capacity”. Also, “it could be defined through three key requirements: **a)** they have an operational computer of server quality, **b)** they have a DNS independent addressing system” and **c)** they are able to scope with variable connectivity. Also, as defined in [61]: P2P is a class of applications that takes advantage of resources-storage, cycle, content, human presence-availability at the edges of Internet. Because accessing to these decentralized resources means operating in environment with unstable connectivity and unpredictable IP addresses. P2P nodes must operate outside the DNS system and have significant or total autonomy from central servers [48].

In this paper, we give a generalized state of art on P2P networks; we present different classifications with representative examples, P2P platforms and applications, major security problems and P2P simulators with analysis and discussion.

## 2. Peer to Peer Network Architectures

P2P systems implement a virtual overlay network over the underlying physical network as described on Figure 1. There are several proposed architectures. In this section, we give some descriptions of the most important P2P protocols with a classification based on the underlying architecture. Based on this concept, we can classify them into two categories: structured and unstructured systems.

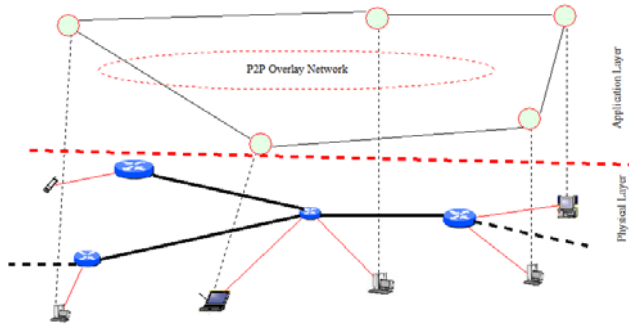


Figure 1: P2P Overlay Network

## 2.1 Unstructured Peer to Peer Networks

Unstructured P2P systems refer to P2P systems with no restriction on data placement in the overlay topology. Some P2P systems also provide the query functionality to locate the files by keyword search. The most significant difference of unstructured P2P systems to the traditional client/server architecture is the high availability of files and network capacity among the peers. Replicated copies of popular files are shared among peers. Instead of downloading from the central server, peers can download the files from other peers inside the network. Obviously, the total network bandwidth for popular file transmissions is undoubtedly greater than the amount most central server systems able to provide. In this sub section, we give functional principles of some unstructured Peer to Peer architectures.

### 2.1.1 Example illustrations

Napster [30] is known as music exchange system. Nodes login to a server and send a list of files that can offer, then issue queries to the server to find which other nodes hold their desired files, and finally download the desired objects directly from the object home. In BitTorrent [24], a peer that wishes to download a certain file joins a group of about 50 other peers that either upload or download this same file. This group is called the **swarm**. The file is split into many small chunks, and these chunks are exchanged by the peers in the swarm. The advantage of this is that peers that download a chunk can upload it to the other peers that do not yet have this chunk, so all chunks of the file are available from multiple peers. This is known as the barter mechanism. Gnutella [2] is a decentralized protocol for distributed search in a flat topology of peers (*servents*). In Gnutella, a querying peer sends a query to all of its neighbour peers, who in turn send the query to all of neighbour, and this spreading broadcast continues until the query reaches a peer that has a file that matches the query, or until a certain predefined maximal number of forwards are reached. If a peer is reached, it sends back a reply containing its address, the size of the file, speed of transfer, etc. The reply traverses the same path as the

query but in a reverse order back to the querying peer. In this way, query is propagated  $N^p$  other peers (*where  $N$  is the number of neighbour peers and  $p$  is the maximum number of forwards 'TT' of the query*). This passing of messages generates much traffic in the networks often leading to congestion and slow response. Freenet [29] is a purely decentralized loosely structured system. It is essentially pools unused disk space in peer computers to create collaborative virtual file system providing file security and publisher anonymity. Freenet provides file-storage service rather than file-sharing service unlike Gnutella. In ECSP [12], peers are grouped into clusters according to their topological proximity, and super-peers are selected from regular peers to act as cluster leaders and service providers. These super-peers are also connected to each other, forming a backbone overlay network operating as a distinct application. The overlay is managed using an application level broadcasting. HPPC [14] (*Hierarchical Projection Pursuit Clustering*) is a clustering architecture; a cluster is characterized by regions of high density separated by regions that are sparse. DV-Flood is [6] organized as a clusters connected through gateways (*super nodes*). Each cluster is represented by a leader using an election/selection algorithm. DV-Flood uses flooding technique for resource retrieval and localization in inter-clusters and intra-clusters. In the first case the flooding is limited by a parameter  $V$  and in the second way, it is limited by a parameter  $V$ . The leaders are used only for routing acceleration, not for data replication or storage.

### 2.1.2 Analysis and Discussion

Unstructured P2P systems can support partial keyword search. These systems depend on blind search techniques, such as flooding and random walk. Hence, the generated volume of query traffic does not scale with the growth in network size. Many research activities are aimed towards improving the routing performance of unstructured P2P systems by adopting hint-based routing strategies. Peers learn from the results of previous routing decisions, and bias future query routing based on this knowledge. Unstructured P2P networks offer a number of important advantages: (1) an unstructured network imposes very small demands on individual nodes, and more specifically it allows nodes to join or leave the network without significantly affecting the system performance. (2) Unstructured networks are appropriate for content-based retrieval (*e.g., keyword searches*) as opposed to object identifier location of structured overlays. (3) Finally, unstructured networks can easily accommodate nodes of varying power. Consequently, they scale to large sizes and they offer more robust performance in the presence of node failures and connection unreliability. According to [27], if scalability concerns were removed from unstructured P2P systems, they might be the preferred

choice for file-sharing and other applications where the following assumptions hold: **(1)** Keyword searching is the common operation, **(2)** most content is typically replicated at a fair fraction of participating sites and **(3)** the node population is highly transient. Table 1 presents some complexities of unstructured P2P networks.

**Tab 1.** Complexities of some unstructured P2P systems

Architecture	Space complexity	Cost lookup
Napster	$n$	$O(1)$
Gnutella	$O(n)$	$O(n)$
Freenet	<i>Hops to Leave</i>	<i>Hops to Leave</i>
DV-Flood	$O(D*V)$	$O(D*V)$

## 2.2 Structured Peer to Peer Networks

Structured P2P networks have emerged mainly in an attempt to address the scalability issues that unstructured systems are faced with. The random search methods adopted by unstructured systems seem to be inherently not scalable [27], and structured systems were proposed, in which the overlay network topology is tightly controlled and files (*or pointers to them*) are placed at precisely specified locations. These systems provide a mapping between the file identifier and location, in the form of a distributed routing table, so that queries can be efficiently routed to the node with the desired file. Structured P2P systems are also referred to as Distributed Hash Tables (DHTs). The evolution of research in DHTs was motivated by the poor scaling properties of unstructured P2P systems.

Thus, a key contribution of DHTs is extreme scalability. As a result, DHTs exhibit some very unique properties that traditional P2P systems lack [49]. Generally, P2P overlay network is characterized by the decisions made on the following six key design aspects [54]:

- 1) Choice of an identifier space, 2) Mapping of resources and peers to the identifier space, 3) Management of the identifier space by the peers, 4) Graph embedding (*structure of the logical network*), 5) Routing strategy and 6) Maintenance strategy.

In this sub section, we give functional principles of some structured P2P architectures.

### 2.2.1 Example illustrations

In Plaxton [31], each node or machine can take on the role of servers (*where objects are stored*), routers (*which forward messages*) and clients (*origins of requests*). Objects and nodes have names independent of their location and semantic properties, in the form of random fixed length bits sequences represented by a common base. Object location in plaxton works as follows: **1)** a servers S1 publishes that it has an objects O1 by routing a

message to the root node of O1. The published process consists of sending toward the root node a message, which contains a mapping  $\langle object-id, server-id \rangle$ . **2)** During object location, a query message destined for object O1 is initially routed towards O1's root. **3)** At each step, if the message encounters a node that contains the location mapping for O1, it is immediately redirected to the server containing the object O1. Otherwise, the message is forward one step closer to the root. If the message reaches the root, it is guaranteed to find a mapping for the location of O1. Chord [5] is a decentralized peer to peer lookup service that stores key/value pairs for distributed data items. Given a key, the node responsible for storing the key's value can be determined using a hash function that assigns an identifier to each node and to each key. Each key  $k$  is stored on the first node whose identifier  $id$  is equal or follows  $k$  in the identifier space. DKS [39] stands for Distributed k-ary Search and it was designed after perceiving that many DHT systems are instances of a form of k-ary search. A query arriving at a node is forwarded to the first node in the interval to which the  $id$  of the node belongs. Therefore, a lookup is resolved in  $\log_k(N)$  hops. CAN [3] is a distributed and structured P2P lookup services, each key will be evenly hashed into a point of d-dimensional space, as its identifier, when a node joins, it will randomly select a point of d-dimensional space. Then, it will be responsible for half of regions this point belongs to, and holds all keys who's IDs belongs to this region. Each node will keep its neighbour node ID locally, and routing is then performed by forwarding request to the regions closest to the position of the key. The expected search length is  $O(d\sqrt[d]{N})$  and state information kept locally is  $O(d)$ . FCAN [32] is based on CAN and propose a kind of search scheme in structure P2P that supports semantic based query. The construction of P2P overlay should ensure that the organization of peers in P2P system and the placement of the data objects are consistent with the semantic space that they belong to. ABC [7] is also called Alpha-Beta cluster-based protocol. A cluster of nodes work together to offer efficiency routing, and the size of each cluster can vary between an upper bound (*ALPHA*) and a lower bound (*BETA*). Each node maintaining  $O(\log(n))$  logical links. ABC can achieve each query within  $O\left(\frac{\log(n)}{\log \log n}\right)$  hops in the structured P2P

system, where  $n$  is the total number of nodes in the system. CISS [8] is a collaborative information shared system based DHT, it uses a locality based preserving function (LPF) instead of a hash function. CISS consists of a client and server modules. The client module takes the updates of queries, it routes them to rendezvous peer nodes for processing. The server module stores objects to its repository and processes incoming queries, it returns

matched results to requesting peer nodes. The key idea of RPS [18] is to partition the key space (*or the identifier space*) into multiple non-overlapping search regions and assign a region to a node that receives a query message. Each node, on receiving a query message, is allowed to forward the query to other nodes only within its search region. Furthermore, at each step of query forwarding, the forwarding node partitions its search region into smaller regions, i.e., the search region starts as the whole identifier space and gets partitioned recursively as the query messages are forwarded. The search region information is carried in the query message as a tag. Pastry [4] assumes a circular identifier space and each has a list connecting of  $\frac{L}{2}$  successors and  $\frac{L}{2}$  predecessors known as a leaf set. A node also keeps track of  $M$  nodes that are close according to another metric other than the *id* space like network delay. This set is known as the neighbourhood set and is not using during routing but using for maintaining locality properties. The third type of node state is the main routing table. It contains  $\log_2^b(N)$  rows and  $2^b-1$  columns.  $L$ ,  $M$  and  $b$  are system parameters. Tapestry [1] is a self-organizing routing and object location system. Like pastry it is based on the earlier work of Plaxton. It provides routing of messages directly to the “closest” copy of an object (*or service*) using only point-to-point links between nodes and without centralized resources. Location information is distributed within the routing infrastructure and is used for incrementally forwarding messages from point to point until they reach their destination. This information is repairable “*soft-state*”, its consistency is checked on the fly, and if lost due to failures or destroyed, it is easily rebuilt or refreshed. GTapestry [13] assembles physically neighbouring nodes in the Internet into self-organized groups. The routing mechanism of GTapestry is divided into inter-group routing, used by groups that communicate with others through their leaders, and intra-group routing, used by group members to communicate with each other directly.

Kademlia network [15] partitions the identifier space exactly like pastry. However, the node *ids* are leafs of a binary tree where each node’s position is determined by the shortest unique prefix of its *id*. Each node divides the binary tree into a series of a successively lower subtree that don’t contain the node *id* and keeps at least one contact in each of those subtree. Kademlia does not keep a list of nodes close in the identifier space like the leaf set or the successor list in Chord. However, for every subtree/interval in the identifier space, it keeps  $k$  contacts rather than one contact if possible, and calls a group of no more than  $k$  contacts in a subtree. P4L [20] uses a hierarchical rings for content distribution, each ring is similar to Chord in routing, management. Two neighbouring rings communicate between them using a

node belong to these two rings (*relay node*). The cost lookup in P4L is  $O(\sum_{i=1}^4 n_i)$  where  $n_i$  is the number of

nodes on ring level  $i$  (*with maximum of 256 nodes in each ring*). Palma [16] is location management in mobile environment based on the Tapestry algorithm. Palma architecture is composed of heterogeneous wireless and wired networks connected via a high speed wired backbone network and extended with a number of distributed location servers (LSs). Their LSs are organized into an overlay network to publish location information to each other for storage, and to collaboratively resolve queries. In DPMS [11], advertised patterns are replicated and aggregated by the peers, organized in a lattice like hierarchy. Replication improves availability and resilience to peer failure, and aggregation reduces storage overhead. In DPMS a peer can act as a leaf peer or indexing peer. A leaf peer resides at the bottom level of the indexing hierarchy and advertises its indices (*created from the objects it is willing to share*) to other peers in the system. An indexing peer, on the other hand, stores indices from other peers (*leaf peers or indexing peers*). A peer can join different levels of the indexing hierarchy and can simultaneously act in both the roles. Indexing peers get arranged into a lattice like hierarchy and disseminate index information using repeated aggregation and replication. DPMS uses replication trees for disseminating patterns from leaf peers to a large number of indexing peers. However, such a replication strategy would generate a large volume of advertisement traffic. To overcome this shortcoming, DPMS combines replication with lossy-aggregation, advertisements from different peers are aggregated and propagated to peers in the next level along the aggregation tree. DCFLA [10] is a distributed user profile management scheme using distributed hash table (DHT) based routing protocols. SkipNet [35] is a ring approach based on the SkipList. This last is a sorted linked list that contains supplementary pointers at some nodes that facilitate large jump in the list in order to reduce the search time of a node in the list. The idea is applied to the ring structure, where nodes maintain supplementary pointers in the circle identifiers space. SkipNet facilitates placement of keys based on nameID scheme that allows the key to be stored locally or within a confined administrative domain (*path locality*). In addition, SkipNet also provides path locality by restricting the lookups in the DHT only to domains that may contain the required key. Koorde [33] implements De Bruijn graphs on top of ring architecture. A De Bruijn graph maintains two pointers to each node in the graph, thereby requiring only constant state by node in the ring, specifically, given that each node ID is represented as a set of binary digits, each node is connected to nodes with identifiers  $2m$  and  $2m + 1$  (*where*

$m$  is a decimal value of the node's ID). These operations can be regarded as simple left shift and additions of the given node's ID. Therefore, by succession shifting of bits, lookup time of  $\log(n)$  can be maintained. Panache [17] aggregates popularity information and builds upon other peer-to-peer systems that distribute index information by keyword. Relying on a combination of Bloom filtering, query ordering, and truncated results based on popularity data. Tarzan [19] is a P2P anonymizing network layer. A message initiator chooses a path of peers pseudo-randomly through a restricted topology in a way that adversaries cannot easily influence. Tarzan provides anonymity to both clients and server, without requiring that both participate, it uses NAT to bridge between Tarzan host and obvious internet hosts. Cycloid [9] is a constant-degree P2P architecture, which emulates a cube-connected cycles graph in the routing of lookup requests. Cycloid combines Pastry with cube-connected cycle graphs. In a Cycloid system with  $n = d \cdot 2^d$  nodes at most, each lookup takes  $O(d)$  hops with  $O(1)$  neighbours per node. Cycloid is not necessarily complete; it can have nodes less than  $d \cdot 2^d$  with some void node places. Like Pastry, it employs consistent hashing to map keys to nodes. A node and a key have identifiers that are uniformly distributed in a  $d \cdot 2^d$  identifier space. Viceroy [34] is based on the butterfly graph, like many other systems, it organizes nodes into a circular identifier space and each node has successors and predecessors pointers. Moreover, in N-nodes network, nodes are arranged in  $\log_2(N)$  level numbered from 1 to  $\log_2(N)$ . Each node apart from nodes at level 1 has "up" pointer and every node apart from the nodes at the last level 2 "down" pointers. There is one short and one long "down" pointers. Those three pointers are called the butterfly pointers. All nodes also have pointers to successors and predecessors pointers on the same level. In such way, each node has a total of 7 outgoing pointers.

### 2.2.2 Analysis and Discussion

Structured systems offer a scalable solution for exact-match queries, i.e. queries in which the complete identifier of the requested data object is known (as compared to keyword queries). There are ways to use exact-match queries as a substrate for keyword queries [28]. However, it is not clear how scalable these techniques will be in a distributed environment. The disadvantage of structured systems is that it is hard to maintain the structure required for routing in a very transient node population, in which nodes are joining and leaving at a high rate. Table 2 presents some complexities of structured P2P networks.

Tab 2. Complexities of some structured P2P systems

Architecture	Space complexity	Cost lookup
Chord	$O(\log(n))$	$O(\log(n))$
CAN	$2d$	$O(n^{1/d})$
Tapestry	$O(\log_b(n))$	$O(\log_b(n))$
P4L	$\sum_{i=1}^4 \log(n_i)$	$\sum_{i=1}^4 \log(n_i)$
Pastry	$O(\log_2^b(n))$	$O(\log_2^b(n))$
Viceroy	7	$O(\log(n))$
Koorde	2	$O(\log(n))$
Kademlia	$O(\log(n))$	$O(\log(n))$

## 3. Peer to Peer Network Applications

Since the apparition of P2P network, applications are in a continuously grow, from file sharing to real time applications.

**File Sharing:** content storage and exchange is one of the areas where P2P technology has been most successful. File sharing applications focus on storing and retrieving information from various peers in the network. One of the best known examples of such P2P systems is Emule, KaZAa [26].

**Distributed Computing:** these applications use resources from a member of network computers. The general idea behind these applications is that idle cycles from any computer connected to the network can be used for solving the problem of the other computers that require extra computation. SETI@home [25] is one example of such systems.

**Communication and Collaboration:** collaborative P2P applications aim to allow application level collaboration between users. These applications range from instant messaging and chat, to online games, to shared applications that can be used in business, education and home environments. Groove [36] and Jabber [37] are two examples of such systems.

### Analysis and Discussion

P2P networks are known as a file sharing applications. However, it has several kinds of applications as mentioned above. Each C/S application has emerged to P2P application. P2P networking is not restricted to technology, but covers also social processes with a peer-to-peer dynamic. In such context, social P2P processes are currently emerging throughout society.

## 4. Peer to Peer Network Classifications

According to the degree of decentralization of P2P systems, they are divided on three classes represented on table 3.



**Tab 3.** Degree of decentralization based classification

Degree of decentralization	Examples
Purely decentralized	Gnutella, DHT based architectures
Partially centralized	Kazaa, Morpheus
Hybrid decentralized	Napster

In [21], the authors propose a classification based on the type of application. Table 4 presents illustrate this classification.

**Tab 4.** Application categories based classification

Applications	Examples
Communications and collaboration	[37]
Distributed computation	[25], [55]
Internet Service support	[56], [57]
Database systems	[58], [59]
Content distribution	[60]

In [44], the authors propose a classification based on the particularity of each P2P architectures. Table 5 illustrates this classification.

**Tab 5.** Particularity and resemblance based classification

Taxonomy	Selected references
Search	[40], [41], [43]
Ring	[5], [20]
De Bruijn Graph	[47]
Skip Graph	[45], [46]
Key Words Lookup	[2], [42]
locality	[52], [53]

P2P networks can be also classified on 1) semantic consideration on routing: network with semantic routing and without semantic routing. The former is generally based on flooding or local/distributed index and uses key works as in Gnutella. The later is generally based on DHTs such as Chord, P4l. 2) physical proximity consideration: for real time application, lookup is measured by time not by the hops number. Most of P2P networks don't consider physical proximity. 3) Generation according to lifetime cycle: the first generation such as Napster system, the second generation such as Gnutella, and the third generation such as DHT based networks,

### Analysis and Discussion

Many classifications of P2P networks have been presented on the literature. Generally they are focused on routing strategy, because it is the most important and the most executed operation.

## 5. Peer to Peer Network Security

Distributed implementations create additional challenges for security compared to client-server architecture, security in P2P system aims to ensure that the use of the system does not have unwanted influence to a user or environment where the P2P system operates. Achievement a high level of security in peer-to-peer systems is more difficult than non-peer-to-peer systems [38]. The most important attack types are: 1) Replay Attacks: using a previously recorded or captured message to attack a network or to gain access to somewhere one is not authorized to be (*a form of identity theft*), 2) Malicious Provider: a provider that accepts payment but fails to complete the transaction can be contested, 3) Malicious Consumer: a malicious consumer who fraudulently claims that he did not receive services even though he did is thwarted by the use of certificates. The provider simply provides the certificate to his bank-set when the transaction is complete, 4) Routing Attacks: In such case, message routing will fail with high probability, and the systems fail to provide any services , 5) Denial of Service Attack: a DoS attack is an attempt to prevent legitimate users of a service or network resource from accessing that service or resource, 7) Sybil Attacks: in a peer-to-peer domain without external identifiers, any node can manufacture any number of identities.

### Analysis and Discussion

Research concerning security and trust in P2P systems draws upon the expertise of the distributed computing community as well as the sociology community. Even the best protected organizations, companies or personal users are finding it difficult to effectively shield themselves against all malicious security attacks due the increasing rate with which they appear and spread. Distributed implementations of P2P networks create additional challenges for security compared to client-server architecture, especially for reliability, flexibility and load balancing.

## 6. Peer to Peer Platforms

P2P platforms provide infrastructure to support distributed applications using p2p mechanisms. P2P components used in this context are for instance naming, discovery, communication, security and resource aggregation.

**XtremWeb:** it is a P2P project intended to distribute applications over dynamic resources according to their availability and implements its own security and fault tolerance policies [62]. XtremWeb manages tasks following the coordinator worker paradigm. The

coordinator masters the tasks management process. Workers are distributed volunteer entities which use a part of their CPU time to compute tasks provided by the coordinator. Every worker connection is registered by the coordinator, and it requests task to compute accordingly to its own local policy. The workers download task software and all expected objects, stores them and starts computing. When a task is completed, the worker sends the result back to the coordinator.

**Proactive:** it is a project of ObjectWeb Consortium (*ObjectWeb is an international consortium fostering the development of open-source middleware for cutting-edge applications: e-business, clustering, grid computing, management services ...*) [63]. Proactive is a Java library for parallel, distributed and concurrent computing, also featuring mobility and security in a uniform frame- Work with a reduced set of primitives.

**JXTA:** Project JXTA [22] is an open source effort to formulate and implement a set of standard P2P protocols that allow a programmer to build any loosely coupled P2P system. JXTA consists of six protocols that support core P2P operations, such as peer discovery, organization, identification and messaging. JXTA architecture is divided into three layers where it implements the OSI model:

**1) Applications Layer:** this layer implements applications that are integrated to JXTA. Many applications are included such as P2P instant messaging and file sharing. JXTA applications implement the OSI application layer.

**2) Services Layer:** this layer implements services such as searching and indexing, file sharing, protocol translation, authentication and Public Key Infrastructure (PKI) services, as well as many others. JXTA services implement session, presentation and application layers in the OSI model. **3) Platform Layer (JXTA Core):** this layer implements a minimal set of primitives that are common to P2P networking. Primitives include discovery, transport, creation of peers and peer groups and others. The JXTA core implements transport, network and data link layers in the OSI model.

JXTA defines a series of protocols, and XML message formats, for communication between peers [23]. Peers use these protocols to advertise and discover network resources, discover each other, and to communicate and route messages. There are six JXTA protocols: Peer Discovery Protocol (PDP), Peer Resolver Protocol (PRP), Peer Information Protocol (PIP), Peer Membership Protocol (PMP), Pipe Binding Protocol (PBP) and Endpoint Routing Protocol (ERP).

### Analysis and Discussion

The existing P2P networks don't collaborate together even for the same tasks, each one uses its own lookup

mechanism. P2P platforms are then introduced; their main objective is to unify the user communities, and then enables for example Gnutella users to search on Freenet network. P2P platforms provide infrastructure to support distributed applications using p2p mechanisms.

## 7. Peer to Peer Network Simulators

Many P2P network simulators have been proposed in the literature these last years. The most important are resumed on [64]. DHTSim is a discrete event simulator for structured overlays, specifically DHTs. It is intended as a basis for teaching the implementation of DHT protocols, and as such it does not include much functionality for extracting statistics. P2PSim is a discrete event packet level simulator that can simulate structured overlays only. It contains implementations of six candidate protocols: Chord, Accordion, Koorde, Kelips, Tapestry and Kademia. OverlayWeaver provides functionality for simulating structured overlays only and does not provide any simulation of the underlying network. It is packaged with implementations of Chord, Kademia, Pastry, Tapestry and Koorde.

PlanetSim<sup>1</sup> it is an event-based P2P simulator written in Java, it an object oriented simulation framework for overlay networks and services. PeerSim is designed specifically for epidemic protocols with very high scalability and support for dynamicity. It can be used to simulate both structured and unstructured overlays. GPS is a message level discrete event simulator with a built-in protocol implementation of BitTorrent. It allows for simulation of both structured and unstructured overlays.

Neurogridis a P2P search protocol project that includes a single threaded discrete event simulator, originally designed for comparing the Neurogrid protocol, Freenet and Gnutella protocols. The simulator works on the overlay layer level and can simulate either structured or unstructured protocols. It is packaged with implementations of Gnutella, Freenet and the Neurogrid protocols. It is a single threaded discrete event simulator and it does not simulate the underlying network. Query-Cycle Simulator is a P2P file sharing network simulator that uses the Query- Cycle model. In this model, peers, both good and malicious, form an unstructured P2P network. Narses is a scalable, discrete event, flow based application-level network simulator. It allows for modelling of the network with different levels of accuracy and speed to efficiently simulate large distributed applications.

<sup>1</sup><http://projects-deim.urv.cat/trac/planetsim/wiki/PlanetSim>

## Analysis and Discussion

The existing network simulators such as OPNET or NS-2<sup>1</sup> are already used for new P2P models performance evaluation. However, users are confronted by many difficulties, especially on code writing, because these simulators are not conceived for P2P. Many specialized P2P network simulators are then appeared; they integrated more and more routing protocols for facilitating simulation of new P2P models.

## 8. Conclusion and Perspectives

The limitations of client/server systems become evident in large scale distributed environments. P2P networks can be used for improving communication process, optimizing resources discovery/localization, facilitating distributed information exchange. Peer-to- Peer applications need to discover and locate efficiently the node that provides the requested and targeted service. Many P2P architectures have been proposed in the literature, these architectures do not collaborate together, and then P2P platforms have been appeared. Security is an important issue on P2P networking; several attacks are discovered these last years and the major solutions are inspired from these of wireless networks. The existing simulators such as OPNET or NS-2 are not well adapted to P2P. An important number of specialized simulators are then proposed; generally they are focused on routing performance evaluation. This paper presents a generalized and complete survey on P2P activities. Important features that should be addressed on P2P network are performance, scalability, maintenance, reliability, usability, naming, structuring, routing and locating, resource managing, topology updating. As a future works, we envision classifying the mathematical modelling of P2P networks.

## Acknowledgments

The authors would like to thank Mr L. Khenous and Mrs N. Halfoune from Bejaia University for their comments and suggestions.

## References

- [1] Ben.y. Zhao, John kubiadowich and Anthony D. Joseph. *Tapestry: an infrastructure for fault-tolerant Wide-area location and routing*, Report No UCB/CDS-01-1141, Computer Science Division, University of California, Berkeley. April, 2001.
- [2] Gnutella, <http://www.gnutella.com>.
- [3] Sylvia Ratnasamy, Paul Francis, Mark Handley, Richard Karp and Scott Shenker, *A scalable content addressable network*, ACM SIGCOMM, 2001.

- [4] Antony Rowstron and Peter Druschel, *Pastry: a scalable decentralized object location and routing for large scale peer to peer systems*, In Proceedings of the 18<sup>th</sup> IFIP/ACM international conference on distributed systems platforms (*Middleware 2001*), Heidelberg, Germany, November 2001.
- [5] Ion Stoica, Robert Morris, David Liben-Nowell, David Karger, M.Frans Kaashoek, Frank Dabek and Haris Balakrishnan, *Chord: A Scalable Peer-to-Peer lookup Service for Internet Application*, IEEE/ACM Transactions on networking, Vol 11, No. 1, January 2003.
- [6] Mourad Amad and Ahmed Meddahi, *DV-Flood: An Optimized Flooding and Clustering based Approach for Lookup Acceleration in P2P networks*, in proceedings of the International Wireless Communications and Mobile Computing Conference (*IWCMC 2008*), August 2008.
- [7] Xiang Xu, *ABC: A cluster-based protocol for resource location in peer-to-peer systems*, Journal of parallel and distributed computing, 2005.
- [8] Jinwon Lee, Hyonik Lee, Seungwoo Kang, Su Myeon Kim and Junehwa Song, *CISS: An efficient object clustering framework for DHT-based peer-to-peer applications*, Journal of computer networks, 2006.
- [9] Haiying Shen, Cheng-Zhong Xu and Guihai Chen, *Cycloid: A constant-degree and lookup-efficient P2P overlay network*, Journal of performance evaluation, 2006.
- [10] Bo Xie, Peng Han, Fan Yang, Rui-Min Shen, Hua-Jun Zeng and Zheng Chen, *DCFLA: A distributed collaborative-filtering neighbor-locating algorithm*, Journal of information sciences, 2006.
- [11] Reaz Ahmed and Raouf Boutaba, *Distributed Pattern Matching: a Key to Flexible and Efficient P2P Search*, IEEE Journal on selected areas in communications, 2007.
- [12] Juan Li, Son Vuong, *An Efficient Clustered Architecture for P2P Networks*, In Proceedings of the 18<sup>th</sup> International Conference on Advanced Information Networking and Application (*AINA'04*), 2004.
- [13] Hai Jin, Fei Luo, Qin Zhang, Xiaofei Liao and Hao Zhang, *GTapestry: A Locality-Aware Overlay Network for High Performance*, In Proceedings of the 11<sup>th</sup> IEEE Symposium on Computers and Communications (*ISCC'06*) Computing, 2006.
- [14] Alexei D. Miasnikov, Jayson E. Rome and Robert M. Haralick, *A Hierarchical Projection Pursuit Clustering Algorithm*, In Proceedings of the 17<sup>th</sup> International Conference on Pattern Recognition (*ICPR'04*), 2004.
- [15] Petar Maymounkov and David Mazieres, *Kademlia: A peer-to-peer information system based on the XOR metric*. In Proceedings of IPTPS '01.
- [16] Kaouther Sethom, Hossam Afifi and Guy Pujolle, *Palma: A P2P based Architecture for Location Management*.
- [17] Tim Lu, Shan Sinha and Ajay Sudan, *Panache: A Scalable Distributed Index for Keyword Search*. 2003
- [18] Vladimir Vishnevsky, Alexander Safonov and Mikhail Yakimov, *Scalable Blind Search and Broadcasting in Peer-to-Peer Networks*, In Proceedings of the Sixth IEEE International Conference on Peer-to-Peer Computing (*P2P'06*), 2006.
- [19] Michael J. Freedman and Robert Morris, *Tarzan: A Peer-to-Peer Anonymizing Network Layer*, CCS02, Washington, DC, USA, November 2002.

<sup>1</sup><http://www.isi.edu/nsnam/ns/>

- [20] Mourad Amad and Ahmed Meddahi, *P4L: A four layer P2P model for optimizing resources discovery and localization*, APNOMS 2006, LNCS 4238, pp. 342-351.
- [21] John Risson and Tim Moors, *Survey of research towards robust peer-to-peer networks: Search methods*, journal of Computer Networks, Elsevier, vol. 50, pp. 3485-3521, 2006.
- [22] Project JXTA, <http://www.jxta.org>.
- [23] JXTA v2.0 Protocols Specification, <http://spec.jxta.org/nonav/v1.0/docbook/JXTAProtocols.html>.
- [24] BitTorrent, 2005, <http://www.bittorrent.com/>.
- [25] SETI@home, 2001, [setiathome.ssl.berkeley.edu](http://setiathome.ssl.berkeley.edu).
- [26] KaZaA, 2001, [www.kazaa.com](http://www.kazaa.com), 2001.
- [27] Q.Lv, S.Ratnasamy, and S.Shenker. *Can heterogeneity make gnutella scalable?*, In Proceedings of the 1st International Workshop on Peer-to-Peer Systems (IPTPS '02), MIT Faculty Club, Cambridge, MA, USA, March 2002.
- [28] IH. Witten, A.Moffat, and TC. Bell, *Managing Gigabytes: Compressing and Indexing Documents and Images*, Morgan Kaufman, second edition, 1999.
- [29] I. Clarke, O. Sandberg, B.Wiley and T.W. Hong, *Freenet: A distributed anonymous information storage and retrieval system*, Proc. of ICSI Workshop on Design Issues in Anonymity and Unobservability, 2000.
- [30] Napster, <http://www.napster.com>
- [31] CG. Plaxton, R.Rajaraman and AH. Richa, *Accessing nearby copies of replicated objects in a distributed environment*, In Proceedings of ACM SPAA, ACM, June 1997.
- [32] Jing Wang, Shoubao Yang, Ying Gao and Leitao Guo, *FCAN: A Structured P2P System Based on Content Query*, in Proceedings of the Fifth International Conference on Grid and Cooperative Computing (GCC'06), 2006.
- [33] F. Kaashoek and D. Karger, *Koorde: a simple degree optimal hash table*, in the Second Intl Workshop on Peer-to-Peer Systems IPTPS03, February, 2003.
- [34] D. Malkhi, M. Naor and D. Ratajczak, *Viceroy: a scalable and dynamic emulation of the butterfly*, in Proceedings of the 21st Annual Symposium on Principles of Distributed Computing PODC, pp. 183-192, July, 2002.
- [35] N. Harvey, M.B. Jones, S. Saroiu, M. Theimer and A. Wolman, *SkipNet: a scalable overlay network with practical locality properties*, in Proceedings of the Fourth USENIX Symposium on Internet Technologies and Systems USITS03, March 2003.
- [36] Groove, <http://www.groove.net>.
- [37] Jabber, <http://www.jabber.org>.
- [38] B. Pourebrahimi, K. Bertels and S. Vassiliadis, *A Survey of Peer-to-Peer Networks*, In Proceedings of the 16<sup>th</sup> Annual Workshop on Circuits, Systems and Signal Processing, 2005
- [39] Luc Onana Alima, Sameh El-Ansary, Per Brand and Seif Haridi, *DKS(N; k; f): A Family of Low Communication Scalable and Fault-Tolerant Infrastructures for P2P Applications*, In proceedings of the 3<sup>rd</sup> International Workshop On Global and Peer-To-Peer Computing on Large Scale Distributed Systems (CCGRID 2003), Tokyo, Japan, May 2003.
- [40] H. Balakrishnan, M.F. Kaashoek, D. Karger, R. Morris and I. Stoica, *Looking up data in P2P systems*, Communications of the ACM, 46(2), pp. 43-48, 2003.
- [41] S.-M. Shi, Y. Guangwen, D. Wang, J. Yu, S. Qu and M. Chen. *Making peer-to-peer keyword searching feasible using multi-level partitioning*, in proceedings of the 3<sup>rd</sup> Int'l Workshop on Peer-to-Peer Systems, February, 2004.
- [42] James Salter and Nick Antonopoulos, *An optimized two tiers P2P architecture for contextualized keyword searches*, journal of Future Generation Computer Systems, Elsevier, vol. 23, pp. 241-251, 2007.
- [43] B. Yang and H. Garcia-Molina. *Efficient search in peer-to-peer networks*, in Proceedings of the 22<sup>nd</sup> Int'l Conference on Distributed Computing Systems, July, 2002.
- [44] Stephanos Androutsellis-theotokis and Diodimis Spinellis, *A Survey of Peer-to-Peer Content Distribution Technologies*, ACM Computing Surveys, Vol. 36, No. 4, pp. 335-371, December 2004.
- [45] J. Aspnes and G. Shah, *Skip graphs*, in Proceedings of the 14th Annual ACM/SIAM Symposium on Discrete Algorithms, pp.384-393, 2003.
- [46] N. Harvey, M.B. Jones, S. Saroiu, M. Theimer and A. Wolman, *SkipNet: a scalable overlay network with practical locality properties*, in Proceedings of the Fourth USENIX Symposium on Internet Technologies and Systems USITS'03, March 2003.
- [47] A. Datta, S. Girdzijauskas and K. Aberer, *On de Bruijn routing in distributed hash tables: there and back again*, in Proceedings of the Fourth IEEE Intl Conference on Peer-to-Peer Computing, August, 2004.
- [48] Peer-to-Peer Working Group, *Bidirectional Peer-to-Peer communication with interposing Firewalls and NATs*, White Paper, 2001.
- [49] Project IRIS (*Infrastructure for Resilient Internet Systems*), Web-site: <http://www.projectiris.net/>.
- [50] K. Aberer, A. Datta and M. Hauswirth, *The Quest for Balancing Peer Load in Structured Peer-to-Peer Systems*, Technical Report IC/2003/32, 2003.
- [51] J. Aspnes, J. Kirsch and A. Krishnamurthy, *Load balancing and locality in range-queriable data structures*, in Proceedings of the 23<sup>rd</sup> Annual ACM SIGACTSIGOPS Symposium on Principles of Distributed Computing PODC 2004, July 2004.
- [52] K. Gummadi, R. Gummadi, S. Gribble, S. Ratnasamy, S. Shenker and I. Stoica, *The impact of DHT routing geometry on resilience and proximity*, in Proceedings of 2003 Conference on Applications, Technologies, Architectures and Protocols for Computer Communications, pp. 381-394, 2003.
- [53] F. Dabek, R. Cox, F. Kaashoek and R. Morris, *Vivaldi: a decentralized network coordinate system*, SIGCOMM'04, August 2004.
- [54] Karl Aberer, Luc Onana Alima, Ali Ghodsi, Sarunas Girdzijauskas, Seif Haridi and Manfred Hauswirth, *The essence of P2P: A reference architecture for overlay networks*, in Proceedings of the Fifth IEEE International Conference on Peer-to-Peer Computing (P2P05), 2005.
- [55] Genome@Home, 2003, The genome@home project, web site. <http://genomeathome.stanford.edu/>.
- [56] Dario Pompili, Caterina Scoglio and Luca Lopez, *Multicast algorithms in service overlay networks*, Journal of Computer Communications, Elsevier vol. 31 pp. 489-505, 2008.
- [57] H.Q. Guo, L.H. Ngho, W.C. Wong and J.G. Tan, *DINCast: a hop efficient dynamic multicast infrastructure for P2P computing*, Journal of Future Generation Computer Systems, Elsevier, vol. 21, pp. 361-375, 2005.

- [58] Halevy, A., Ives, Z., Mork, P., and Tatarinov, I, *Piazza: Data management infrastructure for semantic web applications*, In Proceedings of the 12<sup>th</sup> International Conference on World Wide Web. Budapest, Hungary, pp. 556-567, 2003.
- [59] Verena Kantere, Dimitrios Tsoumakos, Timos Sellis and Nick Roussopoulos, *GrouPeer: Dynamic clustering of P2P databases*, Journal of Information Systems, Elsevier, doi:10.1016/j.is.2008.04.002, 2008.
- [60] Waldman, M., Ad, R., and Lf, C, *Publius: A robust, tamper evident censorship-resistant web publishing system*, In Proceedings of the 9<sup>th</sup> USENIX Security Symposium, 2000.
- [61] R. Steinmetz and K. Wehrle, *peer to peer Systems and applications*, (Eds) Springer LNCS 3485, 2006.
- [62] XtremWeb, <http://www.xtremweb.net/>.
- [63] INRIA, <http://www.proactive.inria.fr/>.
- [64] Stephen Naicken, Anirban Basu, Barnaby Livingston and Sethalat Rodhetbhai, *A Survey of Peer-to-Peer Network Simulators*, In Proceedings of The Seventh Annual Postgraduate Symposium, 2006.

**Mourad Amad** received the engineer degree from the National Institute of Computer Science (*INI-Algeria*) in 2003 and the magister degree from the University of Bejaia (*Algeria*) in 2005. Currently, he is a PhD student at the University of Bejaia, Member of laboratory L.A.M.O.S. His research interests include peer to peer networks (*architecture, application, security, VoIP*)

**Ahmed Meddahi** is a member of GET/Telecom Lille I Computer Science and Networks department. He obtained his Master degree from University of Lille (*France*) and Ph.D. from University of EVRY (*France*) and "Institut National des Telecommunications". His main interests are focused on IP signalling performance, "VoIP" quality and "context aware" management, P2P. He is associate member of RS2M research group at INT.

**Professor Djamil Assani** was born in 1956 in Biarritz (*Basque Country, France*). He started his career at the University of Constantine in 1978. He received his Ph.D in 1983 from Kiev State University (*Soviet Union*). He is at the University of Bejaia since its opening in 1983/1984. Director of Research, Head of the Faculty of Science and Engineering Science (1999 - 2000). Director of the LAMOS Laboratory (*Modelling and Optimisation of Systems - http://www.lamos.org*), Scientific Head of the Doctoral Computer School (*since its opening in 2003*), he has taught at several universities (Algiers, Annaba, Rouen, Dijon, Montpellier, Tizi Ouzou, Stif,...). He has published many papers on Markov chains, queueing systems, reliability theory, inventory, risk theory, performance evaluation and their applications in some industrial areas as electrical networks and computer systems. He was the president of the national Mathematical Committee (Algerian Ministry of Higher Education and Scientific Research - 1995 - 2005).

# Human Iris Segmentation for Iris Recognition in Unconstrained Environments

Mahmoud Mahlouji<sup>1</sup> and Ali Noruzi<sup>2</sup>

<sup>1</sup> Department of Electrical and Computer Engineering, Kashan Branch, Islamic Azad University  
Kashan, Iran

<sup>2</sup> Department of Electrical and Computer Engineering, Kashan Branch, Islamic Azad University  
Kashan, Iran

## Abstract

This paper presents a human iris recognition system in unconstrained environments in which an effective method is proposed for localization of iris inner and outer boundaries. In this method, after pre-processing stage, circular Hough transform was utilized for localizing circular area of iris inner and outer boundaries. Also, through applying linear Hough transform, localization of boundaries between upper and lower eyelids occluding iris has been performed. In comparison with available iris segmentation methods, not only has the proposed method a relatively higher precision, but also compares with popular available methods in terms of processing time. Experimental results on images available in CASIA database show that the proposed method has an accuracy rate of 97.50%.

**Keywords:** *Biometric, Hough Transform, Segmentation, Normalization, Iris Tissue Encoding, and Matching.*

## 1. Introduction

Iris recognition is a biometric recognition technology that utilizes pattern recognition techniques on the basis of iris high quality images. Since in comparison with other features utilized in biometric systems, iris patterns are more stable and reliable, iris recognition is known as one of the most outstanding biometric technologies [1]. Iris images could be taken from humans eyes free from such limitations as frontal image acquisition and special illumination circumstances. Daugman's [2] and Wildes' [3] systems are the two earliest and most famous iris recognition systems including all iris recognition stages. In Daugman's algorithm, two circles which are not necessarily concentrated form the pattern. Each circle is defined by three parameters ( $x_0$ ,  $y_0$ ,  $r$ ) in a way that ( $x_0$ ,  $y_0$ ) determines the center of a circle with the radius of  $r$ . An integro-differential operator is used to estimate the

values of the three parameters for each circular boundary and the whole image is searched in relation to the increment of radius  $r$ . In Wildes' system, gradient based Hough transform has been used to localize two iris circular boundaries. This system consists of two stages. At first, a binary map is produced from image edges by a Gaussian filter. Then, the analysis is performed in a circular Hough space in order to estimate the three parameters ( $x_0$ ,  $y_0$ ,  $r$ ) for a circle.

In segmentation step of the algorithm proposed in [4], a set of one-dimensional signals is extracted from iris image using the values of illumination intensity on a set of pupil-centered circular contours which have been localized through use of edge detection techniques. In [5] iris images are projected vertically and horizontally to estimate the center of the iris. Also, this method has been utilized for eyelash segmentation and lightening reflection removal in [6]. The algorithm proposed in [7] predicts the optimization of iris biometric system on a bigger set of data on the basis of Gaussian model obtained from a smaller set of data. Also, an iris recognition system has been proposed in [8] which is used for frontal iris images and for an iris image which is not taken from frontal view. When frontal iris image is not available for a particular individual, in this system the issue is considered through maximizing Hamming distance between the two mentioned images or through minimizing Daugman's integro-differential operator. Next, the image is transformed to a frontal image. An algorithm is presented to find eyelash and eyelids occlusions on iris in a completely close up image similar to Daugman's method in [9]. In 3D environment, this algorithm searches for three parameters as with ( $x$ ,  $y$ ) in center and radius of  $z$ . The remainder of present paper is organized as follows. In section 2, the proposed method of iris recognition is

introduced and next in section 3, experimental results of the proposed method on several databases are presented, and finally in section 4, conclusions are drawn.

## 2. Proposed method for iris recognition

Fig. 1 shows block diagram for a biometric system of iris recognition in unconstrained environments in which each block's function is briefly discussed as follows:

1. Image acquisition: in this stage, a photo is taken from iris.
2. Pre-processing: involving edge detection, contrast adjustment and multiplier.
3. Segmentation: including localization of iris inner and outer boundaries and localization of boundary between iris and eyelids.
4. Normalization: involving transformation from polar to Cartesian coordinates and normalization of iris image.
5. Feature extraction: including noise removal from iris image and generating iris code.
6. Classification and matching: involving comparing and matching of iris code with the codes already saved in database.

Regarding the fact that in an unconstrained environment iris may have occlusions caused by upper or lower eyelids or eyes may roll left and rightwards, as the paper goes on, these blocks are introduced and such issues are solved.

### 2.1 Image acquisition

Taking a photo from iris is the initial stage of an iris-based recognition system. Success of other recognition stages is reliant on the quality of the images taken from iris during image acquisition stage. Images available in CASIA database lack reflections in pupil and iris areas because infrared was used for imaging. Additionally, if visible light is used during imaging for those individuals whose iris is dark, a slight contrast comes to existence between iris and pupil which makes it hard to separate these two areas [10].

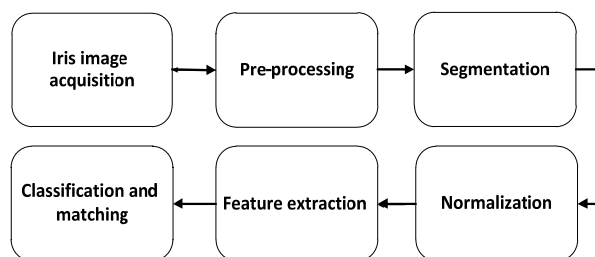


Fig. 1 Block diagram of an iris recognition system.

### 2.2 Pre-processing

Initially, in order to improve and facilitate later processing, a primary processing is performed on iris images. In pre-processing stage, Canny edge detection is used to enhance iris outer boundary that is not recognized well in normal conditions, and a multiplier function is used to enhance Canny iris points, also image contrast adjustment is performed to make its pixels brighter. Fig. 2 shows a sample of an eye image and the results of pre-processing stage performed.

### 2.3 Segmentation

Precise iris image segmentation plays an important role in an iris recognition system since success of the system in upcoming stages is directly dependent on the precision of this stage [16]. The main purpose of segmentation stage is to localize the two iris boundaries namely, inner boundary of iris-pupil and outer one of iris-sclera and to localize eyelids. Fig. 3 shows block diagram of segmentation stage. As it could be seen in this figure, segmentation stage includes three following steps:

1. Localization of iris inner boundary (the boundary between pupil and iris).
2. Localization of iris outer boundary (the limbic border between sclera and iris).
3. Localization of boundary between eyelids and iris.

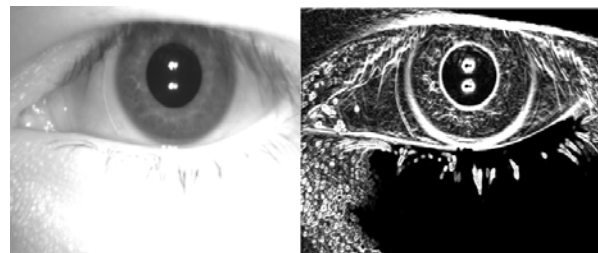


Fig. 2 An eye image from CASIA database and the results of pre-processing performed.

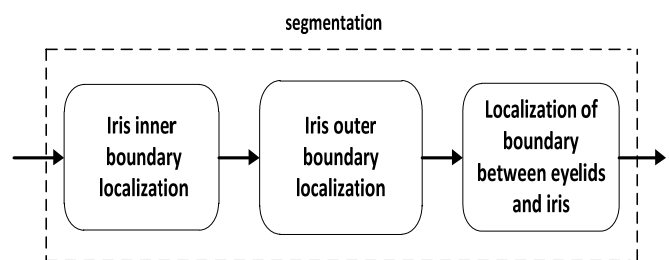


Fig. 3 Block diagram of segmentation stage.

### 2.3.1 Iris inner boundary localization

Regarding that illumination intensity is very different in pupillary inner and outer parts, and pupil is darker compared with iris, the use of Canny edge detection in pre-processing stage results in determining points in iris-pupil boundary. Fig. 2 shows the results of performing Canny edge detection on an eye image as pre-processing output. As it could be observed, pupillary boundary is almost completely detected. After determining edge points, by the use of circular Hough Transform, the center and radius of iris circle are obtained. Fig. 4 shows iris inner boundary which has been achieved via this method for two eye images.

### 2.3.2 Iris outer boundary localization

The most important and challenging stage of segmentation is detecting the boundary of iris and sclera. Firstly, because there is usually no specific boundary in this area and illumination intensity distinction between iris and sclera is very low at the border. Secondly, there are other edge points in eye image in which illumination intensity distinction is much more than that of the boundary of iris and sclera. As a result, edge detection algorithms which are able to detect outer iris edges identify those points as edge. Therefore, in order to detect iris outer boundary, these points have to be identified and eliminated. In this paper, available boundaries are initially enhanced and then extra edge points are identified and eliminated. At the end, through circular Hough transform, outer iris boundary is obtained. In order to enhance iris outer boundary edges, Canny edge detection is performed on eye image in pre-processing stage. By performing such edge detection, a matrix is obtained with the same dimensions as of the image itself which its elements are high in areas where there is a definite boundary and the elements are low in areas where there is no perfectly definite boundary, such as iris outer boundary. Through multiplying of 2.76 in the matrix of pixel values of iris image and intensifying light in eye image, the edges are enhanced. Applying Canny edge detection and multiplying that to the constant value of 2.76 result in better revelation of iris outer boundary edge points. Results of such application on two eye images are shown in Fig. 5.

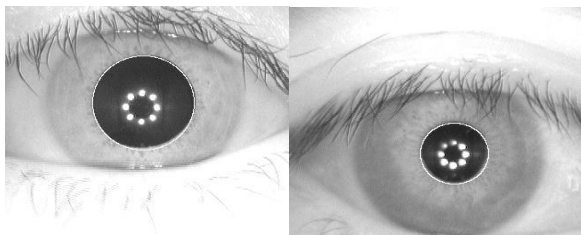


Fig. 4 Iris inner boundary localized for two eye images.

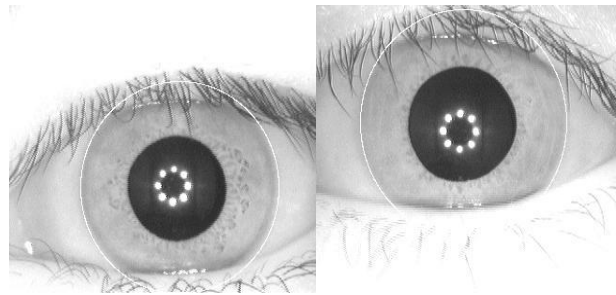


Fig. 5 Iris outer boundary localized for two eye images.

The only issue of this method is sclera boundary not being circular which is the result of angled or sideward imaging and in these cases, some information are lost or clutter comes to existence. In this stage, after identifying iris inner and outer boundaries, the results of these two stages are combined. Fig. 6 shows the results obtained. As it could be seen in this figure, iris inner and outer boundaries are correctly identified in CASIA Iris Image-Interval database.

### 2.3.3 Localization of boundary between eyelids and iris

As could be seen in Fig. 2, one can consider the boundary between eyelids and iris as two lines with first order estimate. To localize them, after detecting edges and by the use of linear Hough transform, the properties of the line could be obtained. To do this, initially eyelids boundary should be detected by using of Canny edge detection. As could be seen in Fig. 2, there are only pupillary edge points between the two eyelids and since pupillary boundary has been already obtained, these points are eliminated. Fig. 7 shows few boundaries localized through this method for some eye images. This method could result in a false outcome only for some images which have too many patterns in iris tissue when the edges of these patterns are detected by Canny edge detection. As they are observable in Fig. 7, the method localizes eyelids with relatively high precision.

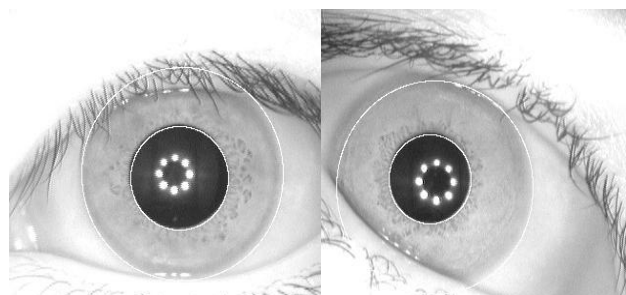


Fig. 6 Iris inner and outer boundaries localized for two eye images.



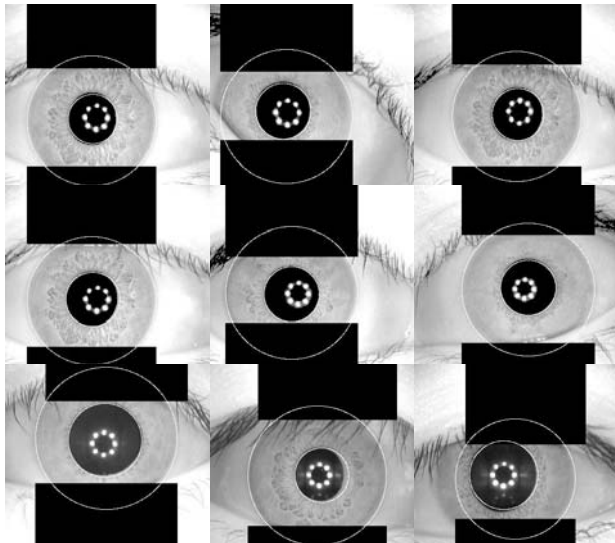


Fig. 7 Boundaries between iris and eyelids localized for some eye images.

They are obvious in Fig. 7, for those eye images in which eyelids are linearly shaped, the boundaries of eyelids and iris are recognized properly but for those images in which eyelids are parabola shaped, this boundary is recognized with slight discrepancy. The accuracy rate of the proposed method for segmentation stage on different databases is presented in Table 1. As the results presented in this table show the method has an accuracy rate of more than 97.6% for iris boundary localization.

## 2.4 Normalization

In normalization stage, an approach based on Daugman's method is used. Fig. 8 shows transforming iris area from polar to Cartesian coordinates. Therefore, iris area is obtained as a normalized strip with regard to iris boundaries and pupillary center. In this paper, iris area is illustrated on a rectangular strip of  $8 \times 512$  [11, 12].

Table 1: Accuracy rate of iris boundary localization for three databases

Database	Accuracy rate in pupil boundary localization (%)	Accuracy rate in sclera boundary localization (%)	Accuracy rate in eyelids boundary localization (%)
CASIA Iris Image (ver. 1.0)	98.13	99	98
CASIA Iris Interval (ver.3.0)	99.73	99.19	98.63
University of Bath	99.18	99.70	97.60

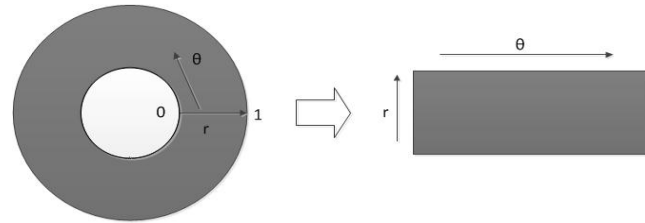


Fig. 8 Transforming polar to Cartesian coordinates.

In order to transform iris area from polar to Cartesian coordinates, 128 pupils-centered perfect circles are chosen starting from iris-pupil boundary and then the pixels located on these circles are mapped into a rectangle. As a result, iris area which looks like a circular strip is converted into a rectangular strip. Since with changes in surrounding illumination level, the size of pupil is adjusted by iris to control the amount of light entering into eyes and it is also possible that individual's distance with a camera could be different, iris is not of the same size in different images. Therefore, choosing these 128 perfect circles normalizes iris in terms of size as well. Then we adjust illumination intensity in segmented iris tissue, i.e. we applied image contrast to bring more clarity into iris tissue. Fig. 9 shows a sample of normalized iris tissue. As the figure shows, all previously mentioned recognition stages have been performed on each image. In initial stage, localization of iris circular inner and outer boundaries, then that of eyelids; later choosing 128 circles on iris area, and eventually transforming polar to Cartesian coordinates has been performed.

## 2.5 Feature extraction and iris encoding

In order to extract features, two-dimensional Gabor Filters are utilized in this paper [13, 14]. Through performing Gabor Filters to the image from different orientations, ultimate feature vector is obtained. In this stage, the dimensions of the feature vector extracted from iris area have to be as small as possible. Regarding high dimensions of the image drawn, Wavelet transform was performed in order to decrease the dimensions in the way that important information existing in tissue can be preserved in spite of downsizing image dimensions [15]. By performing Wavelet transform twice on an image of  $256 \times 512$  already obtained after pre-processing stage, we will have a smaller one of  $16 \times 32$ , and later this image is used to extract features vector [8]. The encoding obtained in this stage with dimensions of  $80 \times 240$  enters the next stage of the system namely matching stage. Regarding that some sections of the area chosen for feature extraction may have occlusions caused by eyelids and eyelashes and since it is possible that, because of error in segmentation stage, some parts of sclera be detected as iris area, it is required that a measure be taken to remove these points from the feature

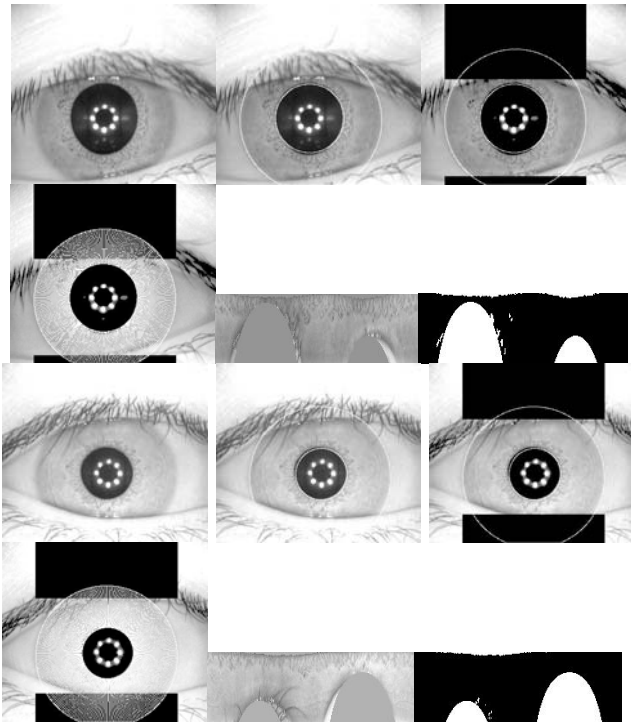


Fig. 9 Transforming iris area into normalized rectangular strip (from left to right: original image, iris inner and outer boundaries, eyelid boundaries, normalization, iris area normalized rectangular strip, eyelids area).

extraction stage. To resolve the latter issue that is caused by error when detecting iris outer boundary, 20% of the lower section of the image is eliminated and to resolve the first issue, points of the image that are placed in this section are eliminated from encoding. To do this, we produce a binary encoding which detects occlusion points. We use this encoding in matching stage and these points are eliminated in that stage. Two outputs are generated in this stage. First output belongs to transformation of iris to iris encoding and another output belongs to transforming iris noises into encodings.

### 2.6 Classification and matching

In the method presented in this paper, Hamming distance criterion is used. If the value of feature vector in a point is equal to the value of other feature vector in that point, digit 1 and if they are not equal, digit zero is allocated to that point and then the values allocated to the pixels are summed up and similarity criterion of the two vectors is attained by finding the best of Hamming distance accordance in following equation:

$$HD = \frac{1}{N} \sum_{i=1}^N (x_i \oplus y_i) \quad (1)$$

In this equation  $N$  is equal to total numbers of feature vector points; and  $x_i, y_i$  is values of two compared feature vectors.

### 3. Experimental results

In this paper, the amount of HD value threshold was supposed to be 0.4. Therefore if two irises are identical, their HD value must be below 0.4 and if two irises are distinctive, their HD value must approximate or exceed 0.4. Efficiency of a biometric system is usually evaluated by taking to account the False Rejection Rate (FRR) and False Acceptance Rate (FAR) and the lower the two error rates; the system is evaluated as more efficient. The FAR means how many people, who introduce you instead of other people, the system accept as an error. The FRR means how many people, who have entrance, allowance, and the system don't known and accept as an error.

In Table 2 the results of the proposed method is presented for two different databases. According to the table results, accuracy rate of proposed method on CASIA database is 97.5% that it is a rather good precision. The reason of being low precision for Bath university database is being very low the difference of illumination intensities in iris and pupil boundary of the database. Also, in Figs. 10 and 11, results of the proposed method for two iris databases are shown. In Table 3, the performance results of several popular algorithm of iris recognition with proposed method on CASIA database images are presented. As seen in the table, the accuracy rate of proposed method is better than Ma algorithm and it is very close to the accuracy rate of Yahya algorithm. It is better mentioned that the reason of very high precision of Daugman's method is very much limitations of this method when it is imaging. Also, linear Hough transform is used for eyelids localization; therefore, the speed of the proposed algorithm is better than the speed of other algorithms such as parabolic Hough transform in the stage of iris localization [16-17].

Table 2: False Rejection Rate (FRR) and False Acceptance Rate (FAR) for two different databases with threshold of 40%

<i>Iris Database</i>	<i>FAR (%)</i>	<i>FRR (%)</i>	<i>Overall system accuracy (%)</i>
CASIA Iris Interval (ver. 3.0)	0.5	2	97.5
University of Bath	4	0	96

Table 3: Efficiency comparison on CASIA database images for popular algorithms

<i>Algorithm</i>	<i>FAR (%)</i>	<i>FRR (%)</i>	<i>Overall system accuracy (%)</i>
Yahya [11]	2.08	0.03	97.89
Daugman [2]	0.09	0.01	99.90
Ma [6]	8.79	0.84	89.37
Proposed method	0.50	2	97.50

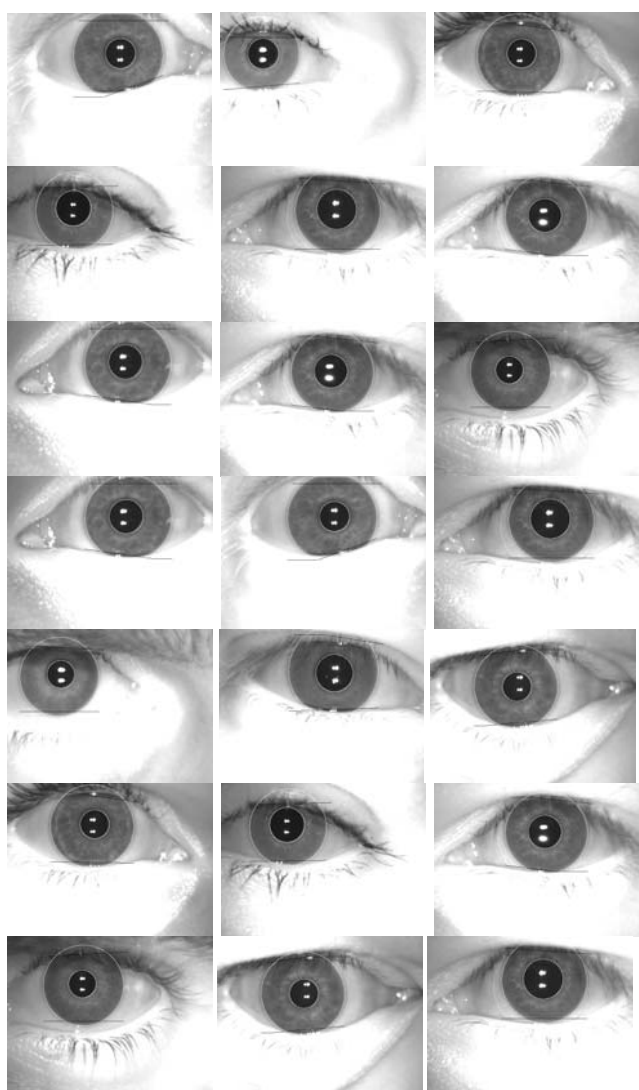


Fig. 10 Iris boundaries localized for some eye images (iris database: University of Bath).

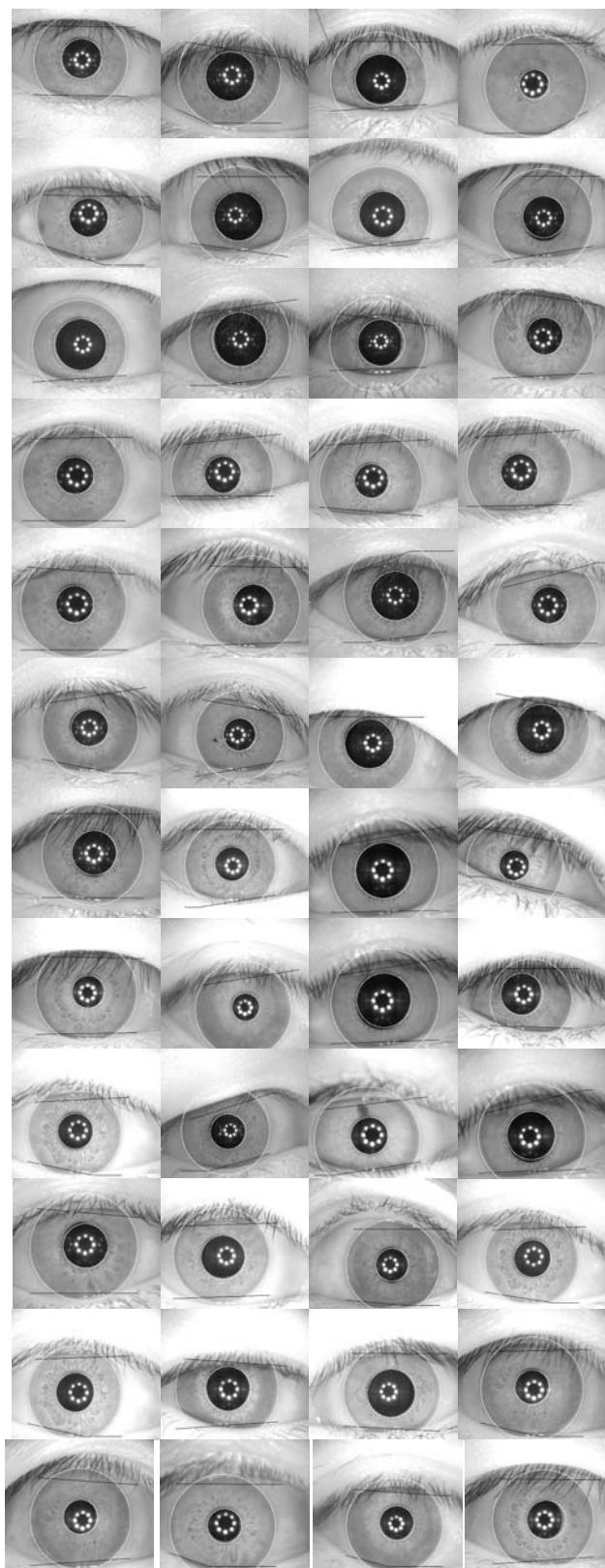


Fig. 11 Iris boundaries localized for some eye images (iris database: CASIA Iris Interval).

## 4. Conclusions

This paper presents an effective method to recognize iris boundaries by performing Canny edge detection and Hough transform. In this method, the boundaries were localized with high precision, and with particular attention to the issue of low variations of illumination intensity in iris outer boundary compared with other sections was achieved a fine accuracy rate for this proposed method. The results of examining the method on CASIA database images indicated the efficiency and high precision of the proposed method that are comparable with other existed methods of identity recognition by using iris images.

## References

- [1] A. K. Jain, A. Ross, and S. Pankanti, "Biometrics: A Tool for Information Security", *IEEE Transactions on Information Forensics and Security*, Vol. 1, No. 2, 2006, pp. 125-143.
- [2] J. Daugman, "New Methods in Iris Recognition", *IEEE Trans. on Systems, Man, and Cybernetics*, Vol. 37, No. 5, 2007, pp. 1167-1175.
- [3] R. Wildes, "Iris Recognition: an Emerging Biometric Technology", *Proceedings of the IEEE*, Vol. 85, No. 9, 1997, pp. 1348-1363.
- [4] W. Boles, and B. Boashash, "A Human Identification Technique Using Images of the Iris and Wavelet Transform", *IEEE Trans. on Signal Processing*, Vol. 46, No. 4, 1998, pp. 1185-1188.
- [5] W. Kong, and D. Zhang, "Accurate Iris Segmentation Based on Novel Reflection and Eyelash Detection Model", in *International Symposium on Intelligent Multimedia, Video and Speech Processing*, 2001, pp. 263-266.
- [6] L. Ma, and T. Tisse, "Personal Recognition Based on Iris Texture Analysis", *IEEE Trans. on PAMI*, Vol. 25, No. 12, 2003, pp. 1519-1533.
- [7] N. Schmid, M. Ketkar, H. Singh, and B. Cukic, "Performance Analysis of Iris Based Identification System the Matching Scores Level", *IEEE Transactions on Information Forensics and Security*, Vol. 1, No. 2, 2006, pp. 154-168.
- [8] V. Dorairaj, A. Schmid, and G. Fahmy, "Performance Evaluation of Iris Based Recognition System Implementing PCA and ICA Encoding Techniques", in *Proceedings of SPIE*, 2005, pp. 51-58.
- [9] C. Fancourt, L. Bogoni, K. Hanna, Y. Guo, and R. Wildes, and N. Takahashi, and U. Jain, "Iris Recognition at a Distance", in *Proceedings of the International Conference on Audio and Video-Based Biometric Person Authentication*, 2005, pp. 1-13.
- [10] "CASIA Iris Image Database", Chinese Academy of Sciences Institute of Automation. <http://www.sinobiometrics.com>
- [11] A. E. Yahya, and M. J. Nordin, "A New Technique for Iris Localization in Iris Recognition System", *Information Technology Journal*, Vol. 7, No. 6, 2008, pp. 924-928.
- [12] L. Masek, "Recognition of Human Iris Patterns for Biometric Identification", *Measurement*, Vol. 32, No. 8, 2003, pp. 1502-1516.
- [13] M. Clark, A. C. Bovik, and W. S. Geisler, "Texture segmentation using Gabor modulation/demodulation", *Pattern Recognition Letters*, Vol. 6, No. 4, 1987, pp. 261-267.
- [14] M. R. Turner, "Texture discrimination by Gabor functions", *Biological Cybernetics*, Vol. 55, No. 2, 1986, pp. 71-82.
- [15] A. Poursaberi, and B. N. Araabi, "An iris recognition system based on Daubechies's wavelet phase", in *Proceedings of the 6th Iranian Conference on Intelligent Systems*, 2004.
- [16] Y. Chen, M. Adjouadi, A. Barreto, N. Rische, and J. Andrian, "A Computational Efficient Iris Extraction Approach in Unconstrained Enviroments", in *BTAS'09 Proceedings of the IEEE International Conference on Biometrics: Theory, Applications and Systems*, 2009, pp. 17-23.
- [17] S. Shah, and A. Ross, "Iris Segmentation Using Geodesic Active Contours", *IEEE Trans. on Information Forensics and Security*, Vol. 4, No. 4, 2009, pp. 824-836.

**Mahmoud Mahlouji** received the B.S. degree in telecommunications engineering from Sharif University of Technology, Tehran, Iran, in 1990, the M.Sc. degree in electronics engineering from Sharif University of Technology, Tehran, Iran, in 1993, and the Ph.D. degree in telecommunications engineering from Science and Research Branch, Islamic Azad University, Tehran, Iran, in 2008. At present he is an assistant professor of the Electrical and Computer Engineering Department, Kashan Branch, Islamic Azad University, Kashan, Iran. His current interests are in image processing, pattern recognition, neural networks, computer vision, and iris recognition.

**Ali Noruzi** received the B.S. degree in computer engineering from Neragh Branch, Islamic Azad University, Neragh, Iran, in 2007, and the M.Sc. degree in computer architecture from Dezful Branch, Islamic Azad University, Dezful, Iran, in 2010. His research interests include image processing, and iris recognition.

# Proposition of Model for CSIRT: Case Study of Telecommunication Company in a Province of Iran

Ali Naseri and Omid Azmoon<sup>1</sup>

<sup>1</sup> Department of Information and Communications Technology, IHU University  
Tehran, Iran

## Abstract

Attack to software, network and system is cause of computer security incidents. Computer Security Incident Response Capability (CSIRC) prevent from attacks by responding, predictive and safety quality management services. Computer Security Incident Response Team (CSIRT) provides these services. In this paper, modeling is proposed for deployment of CSIRT and structure for providing services is described. This model is implemented in a case study in a telecommunication company in a province of Iran and results are expressed.

**Keywords:** CSIRC, CSIRT, Constituency, Computer Emergency Response Team (CERT).

## 1. Introduction

Computer Security Incident Response Team (CSIRT) provides services and supports for preventing, handling and responding of computer security incidents like crash the services, malware distribution, unauthorized access, improper use and combined attacks [1]. CSIRT provide three services that are classified to responding services, predictive services and safety quality management services [2]. For efficient CSIRT operational framework, policymaking frameworks, quality assurance, flexibility and compatibility must be attended [3]. In this paper, modeling is proposed for deployment of CSIRT in awareness, support, disaster relief, and inhibition of the security environment of information exchange, technical services, including investigation of incidents and vulnerabilities, identification and assessment, counseling, education and information network security in the areas inside and outside of companies. The modeling process includes assessment, the steering committee, identify the entities involved in creating a CSIRT, the structure of the CSIRT, identify interactions with domestic and foreign institutions and the process is working and interactions. This model is surveyed in case study in a telecommunication company in a province of Iran and results are expressed. In continue, CSIRC is surveyed and then requirement for establishing of CSIRT is described. Next, model is proposed and case study is expressed and analyses.

## 2. CSIRC

### 2.1. Advantages

Aim of CSIRC is rapid and effective response to computer and network attacks. Therefore, CSIRC response to events by systematic process, evaluate security of infrastructures of hardware and software, help personnel for more effective and quicker services, prevent loss or theft information, use of incident information for future and more powerful protection, develop Standard Operating Procedures (SOP) priorities of organization and more quickly and efficiently respond, deal with legal issues occurred during the events [4].

### 2.2. Types

Different structures are proposed for CSIRT by missions, goals, requirement and services. Teams may be established in different sections like commercial, governmental, military, research and development, IT, ... and have different structures like management, services, organizational model, ... . Therefore, CSIRTs are divided by shape, size and function to internal, national CSIRTs, coordination centers, analysis centers, manufacturer's team and incidents response service. In first step, CSIRT responses to actual security incidents and decreases attack effects. Then statistics and reports of incidents in all fields are synchronized in order to determine security position of the organization and vulnerabilities. Finally, CSIRT play roles like protecting the systems, detecting, diagnosing and analyzing security damages, protecting against destructive activities, cyber security incidents coordination and effective response to computer attacks.

CSIRTs have different organizational models such as the available security teams, centralized model, and distributed model, combined and coordinating model. It is better to use centralized models in small organizations while combined model acts in the best manner for large and disperse covered organizations and centers.

Authority of CSIRT is at three levels, full, shared, no authority. First CSIRTs had no authority and commonly be national or academic type. In local CSIRTs, other levels of authority are needed [5].

### 2.3 Incidents Management

All processes and actions relating to incidents must handle by CSIRT include detect, triage, analyses and response. Incident management includes vast duties from services to functions such as handling vulnerabilities and harmful codes, training, security warnings and ... . Incident response process has different stages include preparation, diagnosis and investigation, limitation, uprooting and restoration, activities and supports after removing the incidents [5] [6].

### 2.4 Implementation

For implementation of CSIRT stakeholders must be identified. Managers must Support, then plan of project are developed. Other steps are: Gathering information, Identification of constituency, definition of mission, funding, determination of level and range of services, determination of organizational model and authority and reporting structure, identification of required resources including staff and equipment and infrastructure, definition of interactions and interfaces, definition of roles and responsibilities, documentation of workflow, development of policies and procedures, development of implementation plan and request feedback, declaration of a formal CSIRT activities, definition of methods for performance evaluation, support for each element having a CSIRT [5] [6].

## 3. The Proposed Model of CSIRT

### 3.1 Needs

Needs and requirements are extracted from sessions by managers and directors and surveyed from experimental documents. Important and usual demands are: establishment of centralized and independent unit, establishment of strategic council, control of computer security incidents in the least time and minimize damage, prevention of accidents by secure systems and prediction of security attacks, risk analysis and evaluation of network status, the training of experts staff and increase skill of computer issues of personnel, attention for security incidents, provide ad hoc and periodic reports, evaluation of security tools and periodic test of permeability and security of applications, provide a copy of the configuration information of the company's equipment and update documents.

### 3.2 Structure

*a. Missions:* CSIRT missions are determined by sessions with directors and managers include: effective control of incident, prevention of incident, improvement of security by identification of risks and threats, network monitoring, preservation of data sources, promoting computer knowledge of user, attention of security of software, development of external stakeholders and provide security services for applicants by receive tariff, support and promote cyber security.

*b. Types:* Types are combination of internal by high priority and provide services for foreign applicants by low priority. CSIRT of Telecommunication Company can be converted to CERT, because of management of networks traffic of provinces and by support from government and security associations.

*c. Sections and Services:* With regard to missions, five sections are necessary: technical operations for incident control and response coordination, education and research for security consultation and training, technical analysis section for analysis incident and damaging codes, support section for development of security tools and maintenance and configuration, customers section for provision agreements service.

*d. Model:* Features of proposed model are: centralized unit of full-time professionals, distribution of number of employees in strategic positions in the organization and the city, addition and combination of reports by central team, high level analysis and provide strategies by central team, implement strategies by members of teams, faster response to incidents by distributed members of team, transfer skills and knowledge to areas of responsibility by distributed members of teams.

*e. Requirements:* Infrastructure and technical requirements are include: communication (like telephone and fax), equipment of CSIRT site (like emergency power system, CCTV, physical security, laptops, printers, digital cameras, audio recording and data storage for keeping records of accident, necessary software and applications), elements of network infrastructure (like valid IP addresses, domain, encrypted email system by PGP with filtering and capability of search for internal and external communications, a secure portal with important information of mission and recommendations and et al., secure configuration of network equipment such as routers for the screening of traffic using ACL, configure a firewall to control access to/from the network, computers and servers), suitable OS depends on skills of team members and type of network performance and costs vary due to the flexibility and generality, applications of incident response team (AIRT) (ability to pursue the incident with support built into the console and comprehensive incident management), technology and communication tools for transfer advice and warnings, virtual private and peer to

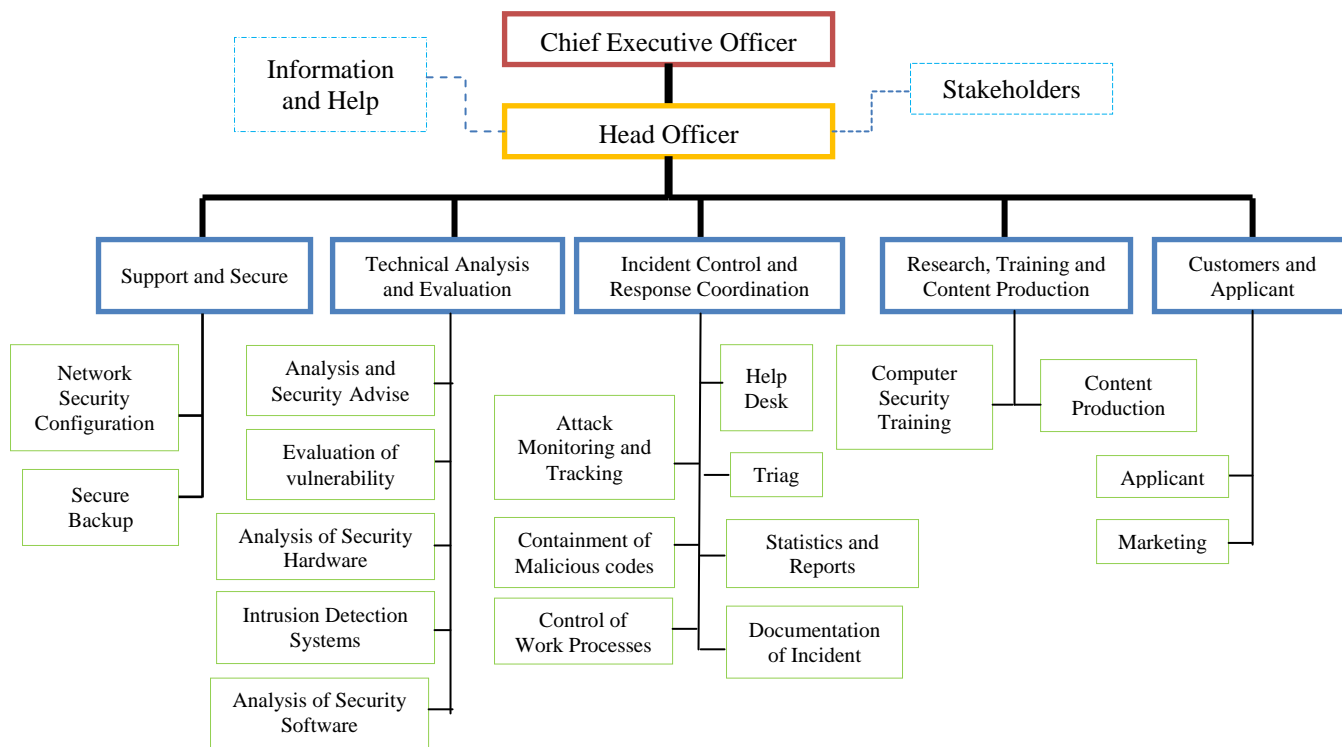


Fig. 1. The Proposed Model of CSIRT Organization

peer networks, legal tools like TCT for Unix, development of backup systems for the last line of defense, development of tools and software (like firewall, Intrusion and Detection Prevention Systems (IDPS), network analyzer, detection and prevent wireless interference, logging events, spam and URL filtering, antivirus and antispayware).

#### 4. Case Study: Telecommunication Company in a Province of Iran

In a province of Iran, managers of Telecommunication Company (TC), department of Information Technology (IT), department of Guard (GU), department of Data (DA) and Financial Section (FS) play role in creation of CSIRT. An internal CSIRT needs interaction with the authorities of physical security, constituency representatives, the organization information technology authorities, telecommunication section, media communication or public relations, business managers, legal consultates, law enforcement, human resources and security risks management section. In table 1, implementation of the proposed model of CSIRT for a telecommunication company in a province of Iran is explained in detail. Figure 2 shows CSIRT interactions of a telecommunication company at time of handling incident.

In diagnosis phase, signs of incident are reported by ICL users/data subscribers (by telephone, fax, email or on-line forms in TCL-CERT portal) to relief desk for users/subscribers in team central section. These signs are combined with reports of TCL-CERT laboratory software and equipment (such as warnings and the recorded incidents)/ status of operators and data department monitoring systems information (such as transfer and switch systems stoppage) and information technology department (such as applied programs incidents, servers and components of the network) /information obtained from public observations and security information and sent to triage phase after recoding in TCL-CERT database.

In triage phase, in case that the signs don't indicate vulnerability or computer security incident, they are blocked (such as failure or noise of the subscribers cable communication or slow public traffic), otherwise, all gathered reports are classified and prioritized and sent to incident /vulnerability analysis phase after recording in TCL-CERT database.

Analysis phase is performed by incident/vulnerability analysis experts in assessment and analysis section in TCL-CERT laboratory. Determining type and estimating intensity of the incident and prioritization on the basis of the damaged sources sensitivity and its potential effects are the most important factors which are performed in this

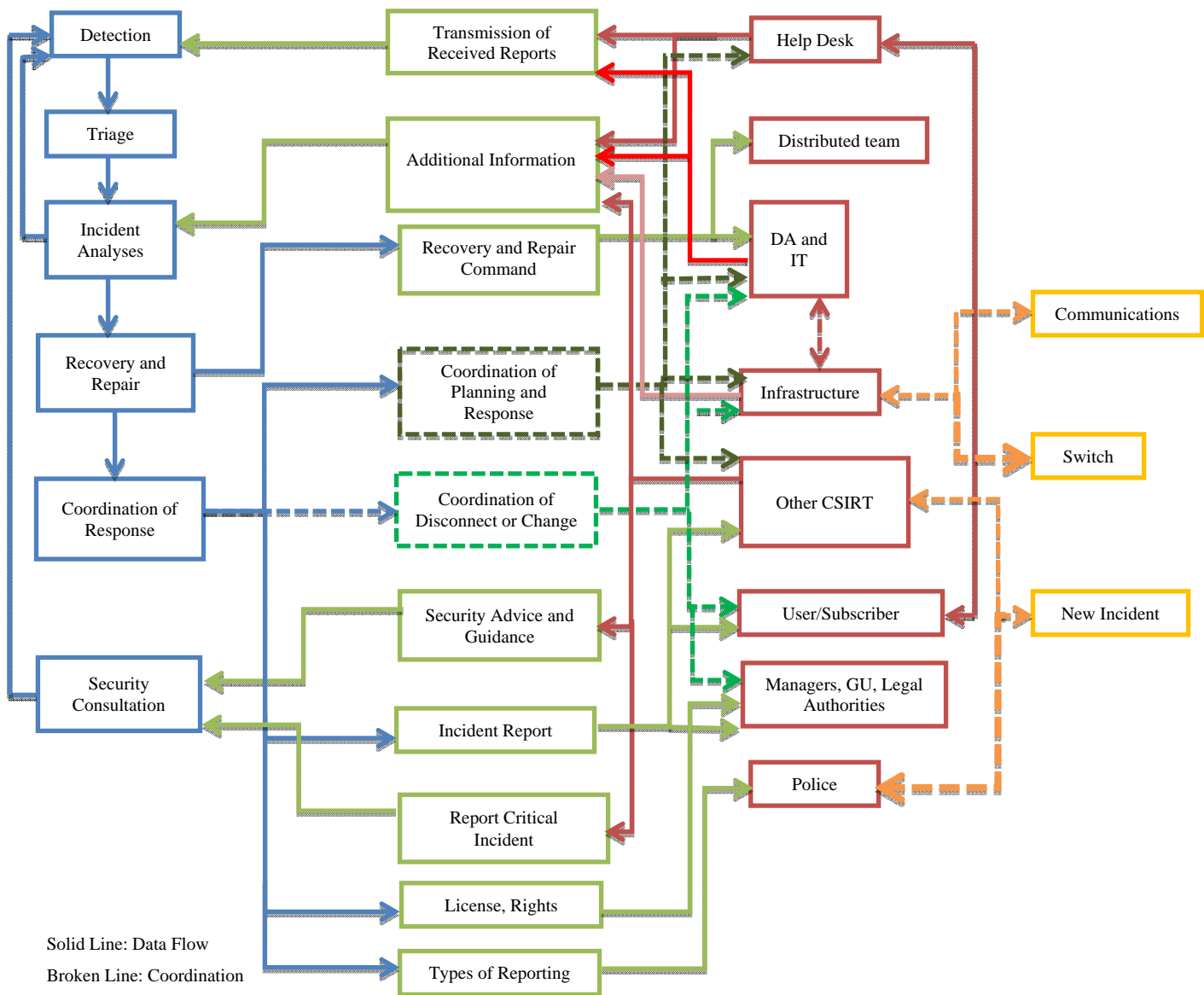


Figure 2: Flowchart of CSIRT in a Province of Iran

phase. If necessary, complementary information is obtained from relief desk of users or subscribers /data and information technology departments/infrastructural NOC/Maher departments or other CSIRTs of the province telecommunication colonies.

Response coordination section should coordinate with directors and authorities of a Telecommunication Company (for internal incidents) /common representative (for external incident), data and information technology departments /infrastructural NOC (if necessary) in case that execution of the response requires stoppage of the user /subscriber or changes in the related communication. Response is executed by distributed team members / data and information technology departments. Incident report is sent to contaminated user/ subscriber for information and in case that the incident requires legal follow-up, it is

referred to legal unit of Telecommunication Company / Justice Administration.

This model is implemented and survey for 6 month and results are indicated in Table 2. Total member of CSIRT is 112 include 24 government employees, 6 responder, 30 coordinator, 12 leader, 30 member and 10 for other works. Reports are added and statistics extracted from them, 437 cases are happened, response to satisfaction is asked and categorized in 4 levels, percentage of satisfaction is summarized in Table 2. Response to importance of problem is asked too and categorized in 4 levels, percentage of importance is summarized in Table 2.



Table 1: Implementation of the Proposed Model by Telecommunication Company in a Province of Iran

Description	Department in TC	Interactions of Internal CSIRT
Telecom CSIRT, Entire company including applicants from the security services. Universities, Registration, Documentation and property records, Banks.	Internal TC: IT, DA, GU External TC: IT of ministries and universities and companies	Constituency Representatives
Provide infrastructures (transmission systems, switches, routers, etc.) Assignment of network services (Internet, Intranet, MPLS, PTMP, etc.) to subscribers (government and private) Management of Information Technology (Web, DB applications, anti-virus, firewalls, etc.)	DA, IT	IT
Provide telecommunications links	Telecom Infrastructure	Telecommunication
Report of incident Training Alert	Public Relations, Communications Portal	Media Communication and Public Relations
Marketing Financial support	Financial and Economic Managers	Business Managers
Rulings of computer crimes	Department of Justice	Legal Consulates
Inspection and follow-up enforcement by the police	Police, Protection Unit	Law Enforcement
Report of incident Presenting symptoms and progressive and provide additional information during the response	Employees, Users	Human Resources
Assessment of threat Identification of financial sector assets Determination of risk	Security Administration, IT, FS	Security Risk Management

Table 2: Statistical Results from Reports

Satisfaction Level	4 (High)	3	2	1 (Low)
Problems (437 cases)	203 or 46.45%	228 or 52.17%	5 or 1.14%	1 or 0.23%
Importance Level	4 (High)	3	2	1 (Low)
Problems (437 cases)	289 or 66.13%	145 or 33.18%	3 or 0.69%	0 or 0%

## 5. Conclusions

Computer Security Incident Response Team (CSIRT) prevent from attacks to software, network and system by responding and management services. In this paper, model for deployment of CSIRT is proposed and structure for providing services is described and implemented in a case study in a telecommunication company in a province of Iran and results are expressed.

## References

- [1] C.Alberts, A.Dorofee, G.Killcrece, R.Ruefle, M.Zajicek, "Defining Incident Management Processes for CSIRTs: A Work in Progress", CMU/SEI, 2004.
- [2] J.Bailey, "Integrating a Leadership and Team Building Module in Community Emergency Response Team Training", Arkansas Tech University, 2009.
- [3] G.Killcrece, R.Ruefle, "Creating and Managing Computer Security Incident Handling Teams (CSIRTs)", Software Engineering Institute, Carnegie Mellon University, 2008.
- [4] K.Scarfone, T.Grance, K.Masone, "Computer Security Incident Handling Guide", U.S. Department of Commerce, National Institute of Standards and Technology NIST SP 800-61, 2008.
- [5] M.West-Brown, D.Stikvoort, K.Kossakowski, G.Killcrece, R.Ruefle, M.Zajicek, "Handbook for Computer Security Incident Response Teams (CSIRTs)", Software Engineering Institute, Carnegie Mellon University, 2003.
- [6] G.Killcrece, K.Kossakowski, R.Ruefle, M.Zajicek, "State of Practice of Computer Security Incident Response Teams (CSIRTs)", Software Engineering Institute, Carnegie Mellon University, 2003.

**Ali Naseri** is Professor of Department of Information and Communications Technology, and Chair for Faculty of Information Processing of IHU University. He has twenty years research experience in processing and computing fields and radar processing. Dr Naseri is also an advisor of Information and Communications Technology Ministry of Iran.



**Omid Azmoon** is Master engineer of Department of Information and Communications Technology of IHU University. He has ten years research experience. His researches have focused on data fusion in radars network. He currently works in Iran Communications Industries.



# Protein sequence for clustering DNA based on Artificial Neural Networks

Gamal. F. Elhadi<sup>1</sup>, R. M. Farouk<sup>2</sup> and Abdalhakeem. T. Issa<sup>3</sup>

<sup>1</sup> Computer Science Department, Faculty of Computers and Information's,  
Menofia University, Menofia, Egypt.

<sup>2</sup> Department of Mathematics, Faculty of Science, Zagazig University,  
Zagazig, Egypt.

<sup>3</sup> Department of Computer Engineering, DCC, Shaqra University, KSA

## Abstract

DNA is a nucleic acid that contains the genetic instructions used in the development and functioning of all known living organisms and some viruses. Clustering is a process that groups a set of objects into clusters so that the similarity among objects in the same cluster is high, while that among the objects in different clusters is low. In this paper, we proposed an approach for clustering DNA sequences using Self-Organizing Map (SOM) algorithm and Protein Sequence. The main objective is to analyze biological data and to bunch DNA to many clusters more easily and efficiently. We use the proposed approach to analyze both large and small amount of input DNA sequences. The results show that the similarity of the sequences does not depend on the amount of input sequences. Our approach depends on evaluating the degree of the DNA sequences similarity using the hierarchal representation Dendrogram. Representing large amount of data using hierarchal tree gives the ability to compare large sequences efficiently

**Keywords:** *DNA Sequences, Protein Sequences, ANN, Clustering*

## 1. Introduction

Most biomacromolecules, such as proteins and nucleic acids, occur in preferred conformations. Examples include n-helices and r-sheets in proteins in nucleic acids. These preferred conformations are the basic keys for structural stability and biological activity of the molecules. Therefore, biological importance to understand the relation between a preferred conformation and the responsible structural parameters. Which structural parameters can be used to study this relation. Possible candidates in the case of nucleic acids are helical parameters, such as, roll, twist and rise [10].

A critical problem in bio-data analysis is to classify bio-sequences or structures based on their critical features and functions. For instance, gene sequences isolated from diseased and healthy tissues can be compared to identify critical differences between the two classes of genes [2]. Such features can be used for classifying bio-data and predicting behaviors. Some approaches have been developed for bio-data classification. For example, one can first retrieve the gene sequences from the two tissue classes and then find and compare the frequently occurring patterns of each class. Usually, sequences occurring more frequently in the diseased samples than in the healthy samples indicate the genetic factors of the disease: those occurring only more frequently in the healthy samples might indicate mechanisms that protect the body from the disease. Similar analysis can be performed on microarray data and protein data to identify similar and dissimilar patterns [7,8].

This work assumes that there is an unknown mapping called clustering structure that assigns a class label to each observation, and the goal of cluster analysis is to estimate this clustering structure, that is, to estimate the number of clusters and cluster assignments. In traditional cluster analysis, it is assumed that such unknown mapping is unique. However, since the observations may cluster in more than one way depending on the variables used, it is natural to permit the existence of more than one clustering structure [12,13,17,23]. This generalized clustering problem of estimating multiple clustering structures is the focus of this paper. The contribution of this paper is to propose a new approach to cluster large DNA data set more efficiently. The proposed system enables researchers to analyze biological data in ease and rapidly using SOM is a specific kind of two-layer Artificial Neural Network

(ANN) can be illustrate by Fig .1, this work proposes an algorithm for finding multiple clustering structures involving clustering variables and observations. The number of clustering structures is determined by the number of variable clusters. The dissimilarity measure for clustering variables is based on nearest-neighbor graphs. The observations are clustered using weighted distances with weights determined by the clusters of the variables.

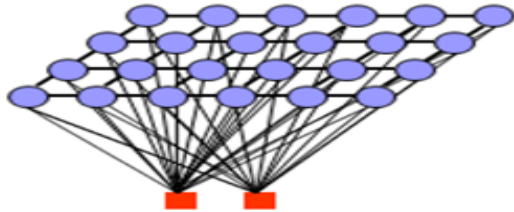


Fig.1 Input  $x_1$  &  $x_2$  SOM layers

The rest of this paper is organized as follows: in sec.2; we illustrate the previous work, in sec3, we discuss our problem methodology, sec.4 include SOM and system analysis algorithm, result and conclusion is given and in sec. 5 we present the future work in sec .6.

## 2. Previous Work

The Basic Local Alignment Search Tool (BLAST) is typically the first Bioinformatics tool that biologists use when examining a new DNA sequence [5,18]. BLAST compares the new sequence to all sequences in the database to find the most similar known sequences [4, 14]. BLAST use Clusters of Orthologous Groups (COGs) [1] tool to compare DNA sequences encoded in 43 complete genomes, representing 30 major phylogenetic lineages. Each COGs consists of groups of paralogs from at least 3 lineages and thus corresponds to an ancient conserved domain. It also uses VAST search to structure similarity search service [4]. It compares 3-D coordinates of a newly determined protein structure to those in the MMDB/PDB database [6].

For the sake of completeness of our investigations, let us offer some introductory notes about the underlying DNA processing. Basically, DNA consists of polymer chains, usually refereed to as DNA strands. This chain is composed of nucleotides, and nucleotides may differ only in their bases. There are four bases which are A (adenine), G (guanine), C (cytosine) and T (thymine).

The familiar double helix of DNA arises by the bonding of two separate strands known as Watson-Crick complementarily, which comes in the formation of such double strands. These four bases are bounded: A always

bonds with T and G with C. Fig. 2 shows DNA bases in a double helix form.

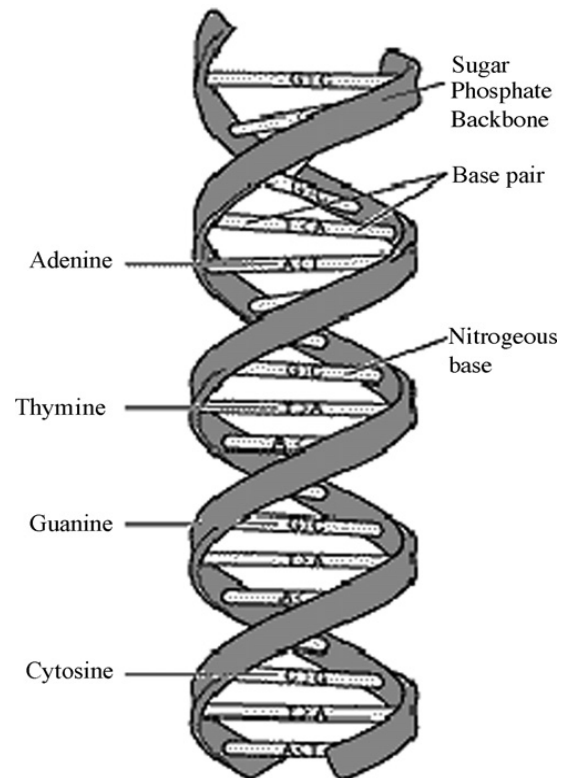


Fig. 2. DNA in a double helix form

Characterizing the DNA-binding specificities of transcription factors is a key problem in computational biology that has been addressed by multiple algorithms. These usually take as input sequences that are putatively bound by the same factor and output one or more DNA motifs [20]. A common practice is to apply several such algorithms simultaneously to improve coverage at the price of redundancy. In interpreting such results, two tasks are crucial: *clustering* of redundant motifs, and attributing the motifs to transcription factors by more DNA motifs. In interpreting such results, two tasks are involving motif comparison. Microarray technology has made it possible to simultaneously measure the expression levels of large numbers of genes in a short time. Gene expression data is information rich; however, extensive data mining is required to identify the patterns that characterize the underlying mechanisms of action. Clustering is an important tool for finding groups of genes with similar expression patterns in microarray data analysis. Hard clustering approaches assign each gene exactly to one cluster, are poorly suited to the analysis of microarray datasets because in such datasets the clusters of genes frequently overlap [10, 11].

### 3. Problem Methodology

In this paper, we propose an approach depends on microarray analysis approach, and the most suitable approach recently used in bioinformatics. We have used SOM approach for clustering DNA, which is very simple, easy to understand and efficient. Matlab tools are used to implement our proposed approach. Clustering DNA sequence is performed using SOM algorithm by comparing DNA sequences based on bases found in the sequences. Then DNA sequence is converted into protein sequence by using the genetic code and save this protein sequence in a database. In the next step clustering is performed based on protein sequence. After getting DNA sequence from the user or from a database, we have converted it into protein sequence by using the standard genetic code and then performing clustering on protein sequence. In addition, we will also perform similarity recognition between DNA sequences and plot a figure of the dendrogram illustrating this similarity. A dendrogram is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering. Dendrograms are often used in computational biology to illustrate the clustering of genes. The proposed approach enables the user to insert to, update or open the database. The framework of our proposed approach is shown in Fig.3.

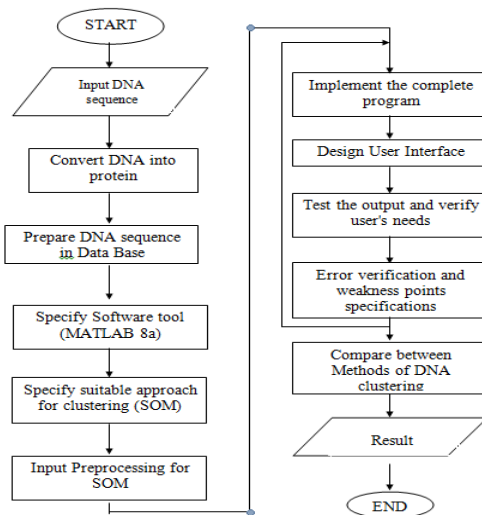


Fig.3. framework of proposed approach

#### 3.1 Self-Organizing Maps

A self-organizing map (SOM) as an explicit brand of Artificial Neural Network (ANN), that is trained using unsupervised learning to produce a low-dimensional (typically two-dimensional), discretized representation of the input space of the training samples, called a map. Self-

organizing maps are different than other artificial neural networks in the sense that they use a neighborhood function to preserve the topological properties of the input space. This makes SOM useful for visualizing low-dimensional views of high-dimensional data, akin to multidimensional scaling. The model was first described as an artificial neural network by Kohonen map [15]. Like most artificial neural networks, SOMs operate in two modes: training and mapping. Training builds the map using input examples. It is a competitive process, also called vector quantization. Mapping automatically classifies a new input vector. A self-organizing map consists of components called nodes or neurons. Associated with each node is a weight vector of the same dimension as the input data vectors and a position in the map space, the usual arrangement of nodes is a regular spacing in a hexagonal or rectangular grid see Fig.4.

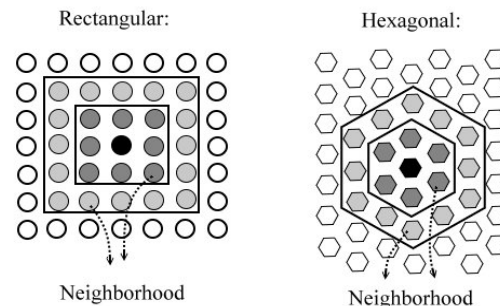


Fig.4. The neurons of the map can be arranged either on a rectangular or a hexagonal map

Self-organizing feature maps (SOFM) learn to classify input vectors according to how they are grouped in the input space. They differ from competitive layers in that neighboring neurons in the self-organizing map learn to recognize neighboring sections of the input space. Thus, self-organizing maps learn both the distribution (as do competitive layers) and topology of the input vectors they are trained on [16].

#### 3.2 Conversion DNA Sequence into Protein Sequence

This section includes the technique used to analysis the proposed system and SOM clustering algorithm. Our proposed approach divides the input into a number of clusters, the value of the clusters must be less than the number of the input sequences, as it is not logical to divide inputs into a number of clusters larger than the number of the inputs. So, by logic if the number of the inputs and the number of clusters are the same, each input will be in a separate cluster and this is of course meaningless [19, 21, 22]. In our proposed approach we accept data from database for DNA sequence [9] and converting each DNA sequence to protein sequence by searching for the starting code (ATG) from DNA sequence then converting each triple of DNA sequence

base (A, C, G and T) to amino acids (using genetic code stored in amino acid database in our proposed approach) and we repeat that for each triple of the sequence bases until we reach the one of the ending codes (TAA, TGA, TAG), these amino acids form a protein sequence [15]. One major problem of SOM is that it requires a value for each dimension of each member of samples in order to generate a map. As we use two-dimensional map, we have to convert each protein sequence into two numbers. These two numbers must be chosen to distinguish each sequence from any others; we proposed a formula for this conversion as follows:

$$x_j = \sum_{i=1}^M A_i(X_i, Y_i) / \sum_{i=1}^M L_i \quad (1)$$

Where  $A_i$  denotes to the ASCII code of each character (A, C, G and T), and  $(X_i, Y_i)$  the position of each character in the sequence,  $L_i$  the length sequence of protein sequence.

#### 4. SOM Clustering Algorithm

After completing preprocessing function and getting the two input numbers, let the map of size  $M$  by  $M$  and the weight vector of neuron  $i$  is  $m_i$ , then we can apply the SOM clustering algorithm as the following steps:

*Step 1:* Initialize all weight vectors  $m_i(0)$  randomly or systematically.

*Step 2:* A vector  $x_j$  is randomly chosen from the training data, then, compute the Euclidean distance  $d_i$  between  $x_j$  and neuron  $i$

$$d_i = \|x - m_i(t)\|, \quad 1 \leq i \leq M^2 \quad (2)$$

*Step 3:* Find the best matching neuron (winning node)  $c$ .

$$d_c = \|x - m_c(t)\| = \min\{\|x - m_i(t)\|\}, \quad \forall i \quad (3)$$

*Step 4:* Update the weight vectors of the winning node  $c$  and its neighborhood as follows.

$$m_i(t+1) = m_i(t) + \alpha(t) h_{c,i}(t) [x - m_i(t)] \quad (4)$$

Where  $0 \leq \alpha(t) \leq 1$  is an adaptive function which decreases with time, the  $h_{c,i}(t)$  is a neighborhood kernel centered at the winning node  $c$ , which decreases with time and the distance between neurons  $c$  and  $i$  in the topology map.

$$h_{c,i}(t) = \exp\left(-\|r_c - r_i\|^2 / \sigma^2(t)\right) \quad (5)$$

Where  $r_c$  and  $r_i$  are the coordinates of neurons  $c$  and  $i$ ,  $\sigma(t)$  is a suitable decreasing function of time,

*Step 5:* iterate the Step 2-4 until the sufficiently accurate map is acquired.

Our program allows the user to enter any sequence and perform similarity to all other inputs according to this sequence. Also the program plot a dendrogram graph to show the similarity tree according to specified sequence.

The distance Euclidian is the step to find the similarity between every pair of objects in the data set. To find the similarity we calculate the Euclidean distance between objects using the (*pdist*) function. Given an  $m$ -by- $n$  data matrix  $X$ , which is treated as ( $1$ -by- $n$ ) row vectors  $x_1, x_2, x_3, \dots, x_n$  the Euclidean distance between the vector  $x_r$  and  $x_s$  are defined from equation (6),

$$(pdist)^2 = (x_r - x_s) \quad (6)$$

After we do similarity according to the processing stage will generate a dendrogram graph which shows the similarity between the specified input and all other inputs. The linkage function takes the distance information generated by (*pdist*) and links pairs of objects that are close together into binary clusters (clusters made up of two objects). The linkage function then links these newly formed clusters to other objects to create bigger clusters until all the objects in the original data set are linked together in a hierarchical tree. The dendrogram is a graphical representation of the results of hierarchical cluster analysis. This is a tree-like plot where each step of hierarchical clustering is represented as a fusion of two branches of the tree into a single one [15,16]. The branches represent clusters obtained on each step of hierarchical clustering.

#### 5. Results and Conclusions

The dendrogram function plots this hierarchical tree information as a graph; the numbers along the horizontal axis represent the indices of the objects in the original data set. The links between objects are represented as upside down U-shaped lines. The height of the U indicates the distance between the objects.

Our approach is implemented using a MATLAB program, the input sequences of DNA sequence [3,11] to the program are clustered into groups with similar sequences. Fig.5. show the comparison of performance for clustering DNA sequences by K-Means, SOM and Linkage Algorithm, where exist similarity between this methods especial SOM Algorithm to clustering DNA after converting into Protein Sequence and K-Means Algorithm.

Fig.6. represents five clusters of two input DNA sequences and its graphical of dendrogram representation the hierarchical clustering of the result is shown in figure 5. We can remark that the subs clusters (1 and 4) are similar in inter cluster distance and also groups (2 and 5) have the same distance. Figure.6 show that two inputs represents ten DNA sequences and its dendrogram is shown in figure.7, in this figure it is clear there are some clusters have the same distance such as (1 and 14), (7 and 9). Figure.8, represents 50 two input sequences and its

graphical representation is shown in figure.9, where number of clusters has the same distance such as (4 and 29) and (8 and 12). Figure .10 represents one hundred two input sequences and its graphical representation is shown in figure.11. The same inter cluster distances refer to the similarity of these clusters. Categorizing the similarity is an efficient and fast process to discriminate the DNA sequences. The results show that our proposed approach is effective in small inputs and also in large inputs. Hierarchical representation of data using dendrogram enables to monitor large number of data easily. The results show that the inter distance between the sub clusters does not relate the amount of data. The inter distance may be long for small two input sequence and vice versa.

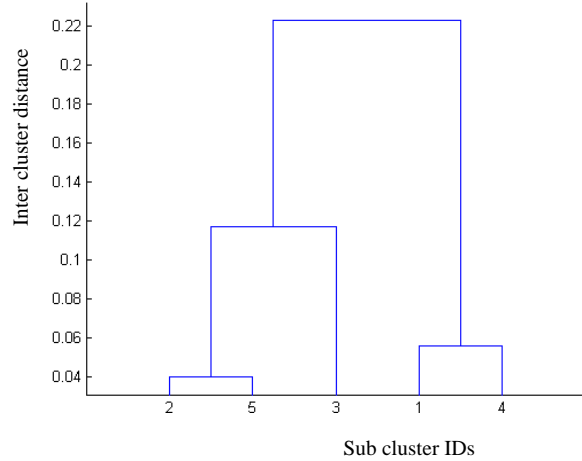


Fig.7. dendrogram of 5 sub clusters

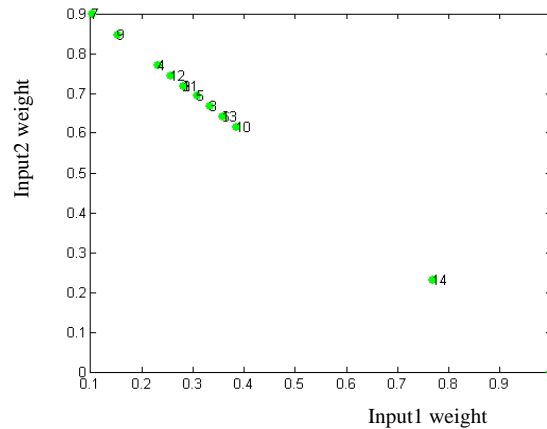


Fig.8. two inputs representation of 10 DNA sequences

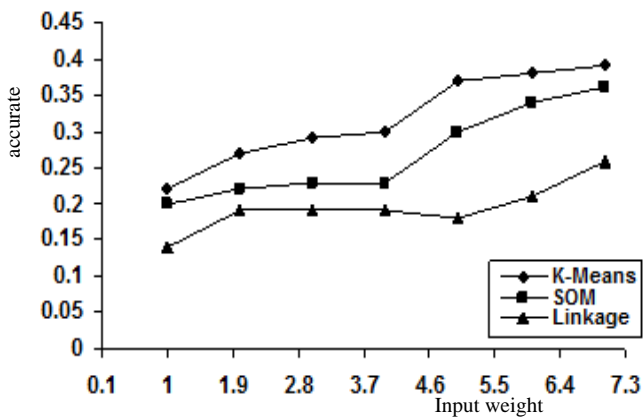


Fig.5. Comparison between K-Means, SOM, and Linkage clustering DNA sequence

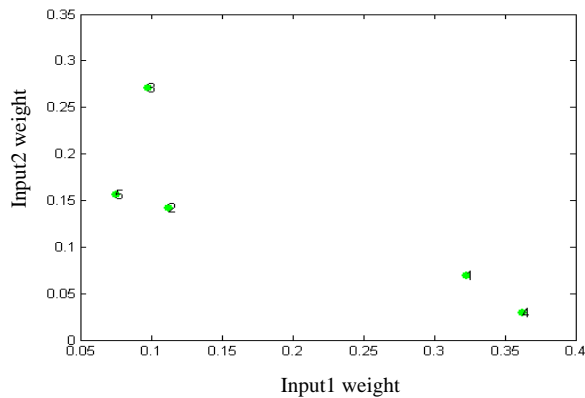


Fig.6. two inputs representation of 5 DNA sequences

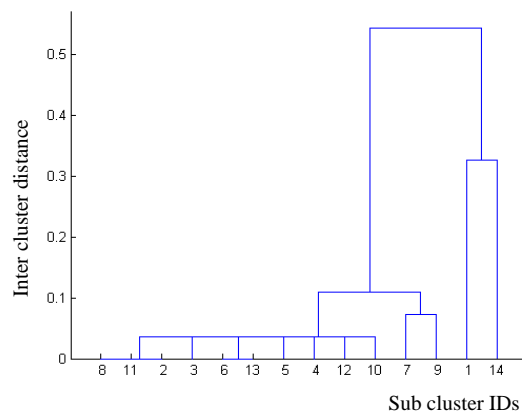


Fig.9. dendrogram of 10 sub clusters

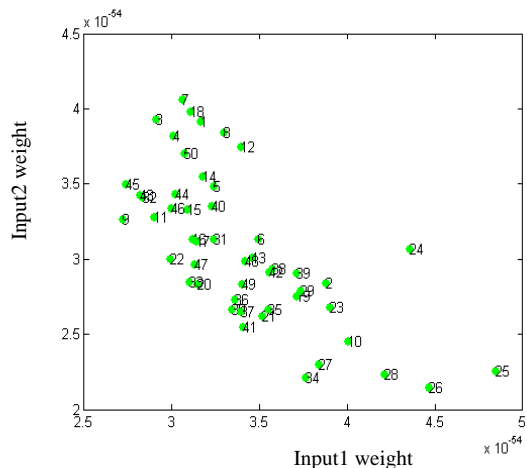


Fig.10. two inputs representation of 50 DNA sequences

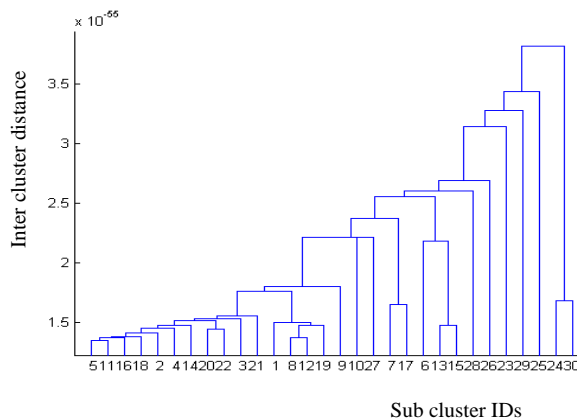


Fig.13. dendrogram of 100 sub clusters

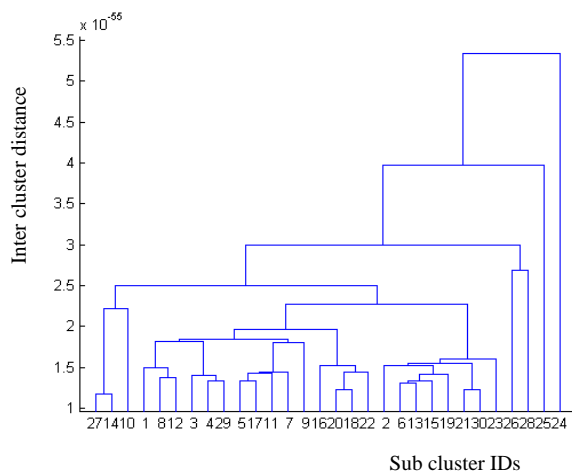


Fig.11. dendrogram of 50 sub clusters

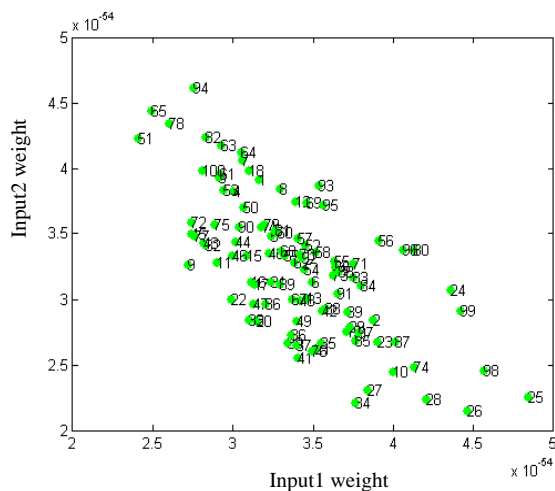


Fig.12. two inputs representation of 100 DNA sequences

This paper proposes an approach uses an efficient SOM way for clustering in DNA sequences, it takes input from database of DNA sequence [3,11]. This helps researches to save notes about the results in a database, or to compare it with previous one to get a new result for discovering diseases or diseases treatment. We divide DNA sequences into a number of cluster groups where the similarity among the sequences in the same cluster is high, while that among the sequences in different clusters is low. And this is very important for many genetic scientists to recognize the similarity or dissimilarity between sequences. The clusters are shown using a graph to be easy to understand. The proposed approach is also able to find similarity between DNA sequences. It finds the similarity between each pair of the inputs. Then, a dendrogram graph is displayed which clearly shows the similarity between the DNA sequences inputs. Moreover, the similarity can be performed according to a particular sequence. In addition, a conversion from DNA sequences into protein sequences can be performed. This operation is important for biologists. After the conversion, the protein sequences are stored in a database. Clustering operation can be performed on these protein sequences.

**References**

- [1] Alberts, Bruce; Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walters (2002). Molecular Biology of the Cell; Fourth Edition.. New York and London: Garland Science. ISBN 0-8153-3218-1.
- [2] Arthur M. Lesk , Introduction to Bioinformatics , ISBN (Pbk)0 19 925196 7,United States by Oxford University Press Inc, 2002.
- [3] Anil K. Jain "Data clustering: 50 years beyond K-means" Pattern Recognition Letters 31 (2010) 651–666.
- [4] Bock, C., S. Reither, T. Mikeska, M. Paulsen, J. Walter, and T. Lengauer. 2005. BiQ Analyzer: visualization and quality control for DNA methylation data from bisulfite sequencing. Bioinformatics 21:4067-4068.

- [5] Beck, S., and V.K. Rakyen. 2008. The methylome: approaches for global DNA methylation profiling. *Trends Genet.* 24:231-237.
- [6] Butler, John M. *Forensic DNA Typing* "Elsevier". pp. 14–15. ISBN 978-0-12-147951-0, 2001.
- [7] Christian Rohde , “New clustering module in BDPC bisulfite sequencing data presentation and compilation web application for DNA methylation analyses” , *BioTechniques*, Vol. 47, No. 3, September 2009, pp. 781–783.
- [8] C. Yu, Q. Liang, C. Yin, R.L. He, S.S.-T. Yau, A novel construction of genome space with biological geometry, *DNA Research* 17 (2010) 155–168.
- [9] Chenglong Yu, Mo Deng, Stephen S.-T. Yau, "DNA sequence comparison by a novel probabilistic method" *Information Sciences* 181 (2011) 1484–1492
- [10] El Hassan, M. A. and Calladine, C. R. (1996) Propellertwisting of base-pairs and the conformational mobility of dinucleotide steps in DNA *Journal of Molecular Biology* 259, 95
- [11] Ehrlich, M., J. Turner, P. Gibbs, L. Lipton, M. Giovanneti, C. Cantor, and D. van den Boom. 2008. Cytosine methylation profiling of cancer cell lines. *Proc. Natl. Acad. Sci. USA* 105:4844-4849.
- [12] Farthing, C.R., G. Ficiz, R.K. Ng, C.F. Chan, S. Andrews, W. Dean, M. Hemberger, and W. Reik. 2008. Global mapping of DNA methylation in mouse promoters reveals epigenetic reprogramming of pluripotency genes. *PLoS Genet.* 4:e1000116.
- [13] Jason T.L. Wang, Mohammed J. Zaki, Hannu T.T. Toivonen and Dennis Shasha (Eds). *Data Mining in Bioinformatics* , ISBN 1852336714, Springer-Verlag London Limited 2005
- [14] Kaufman, L., and P.J. Rousseeuw. 2005. *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, Inc., Hoboken, NJ.
- [15] Kohonen, T. (1989) *Self-organization and associative memory*, 3rd edn. Springer-Verlag, Berlin-Heidelberg.
- [16] Kohonen, T. (1995) *Self-organizing maps*. Springer-Verlag, Heidelberg.
- [17] Ladd-Acosta, C., J. Pevsner, S. Sabunciyen, R.H. Yolken, M.J. Webster, T. Dinkins, P.A. Callinan, J.B. Fan. 2007. DNA methylation signatures within the human brain. *Am. J. Hum. Genet.* 81:1304-1315.
- [18] Meissner, A., T.S. Mikkelsen, H. Gu, M. Wernig, J. Hanna, A. Sivachenko, X. Zhang, B.E. Bernstein. 2008. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* 454:766-770.
- [19] Suzuki, M.M., and A. Bird. DNA methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet.* 9:465-476,
- [20] S. L. Salzberg, D. B. Searls, and S. Kasif, eds., *Computational approaches in Molecular Biology*. Amsterdam: Elsevier Sciences B. V., 1998.
- [21] G. Mecca et al., "A new algorithm for clustering search results", *Data & Knowledge Engineering* (2007) 504–522
- [22] Zhang, Y., et al., “DNA methylation analysis by bisulfite conversion, cloning, and sequencing of individual clones.”, *Methods Mol. Biol.* 2009. 507:177-187.
- [23] Z. Francis, S. Incerti, R. Capra, B. Mascialino, G. Montarou, V. Stepan, C. Villagrasa, Molecular scale track structure simulations in liquid water using the Geant4-DNA Monte-Carlo processes, *Appl. Radiat. Isot.* 69 (1) (2011) 220–226.



# Reliable and Efficient Routing Using Adaptive Genetic Algorithm in Packet Switched Networks

Rakesh Kumar<sup>1</sup> and Mahesh Kumar<sup>2</sup>

<sup>1</sup>Department of Computer Science & Applications, Kurukshetra University Kurukshetra, Haryana, India

<sup>2</sup>Research Scholar, Department of Computer Science & Applications, Kurukshetra University Kurukshetra, Haryana, India

## Abstract

To identify the optimal route is a complex task in packet switched network because optimization depends upon a number of parameters. In this paper Genetic Algorithm is used to locate the optimal route. Genetic Algorithm starts with a number of solutions where each solution is represented in the form of chromosome using the permutation encoding scheme. The success of Genetic Algorithm depends upon the number of operators such as selection, mutation and crossover. Needless to say crossover is most innovative. In this paper crossover operators proposed namely 1-point, 2-point, and adaptive, have been customized according to the need of computer network. The fitness of each solution is evaluated in terms of historical reliability factor, node success/failure and delay. The performance of the proposed approach has been compared with Dijkstra Algorithm and improvement has been observed.

**Keywords:** Crossover, Delay, Dijkstra Algorithm, Genetic Algorithm, Reliability, Routing

## 1. Introduction

Telecommunications networks technologies are growing faster and becoming more complex to handle different types of physical traffic. Today computer networks have application in various domains like business, education, research, Industry, e-commerce and many others. Network based applications are becoming more popular and easy to use with the growth of Internet & development of new technologies & standards. In Data communication networks, such as the Internet and the Mobile Ad-hoc Networks, routing is one of the most important areas that have a significant impact on congestion and network's performance [1]-[3]. Research of the routing strategy is becoming the key theoretical topic of the new generation network architecture. The growth of internet based applications such as on line shopping, Video on demand, video conferencing, on line banking, e-ticket booking, stock exchange and other real time applications has generated new requirements for one-to-one and one-to-many reliable and time efficient

communications. An ideal routing algorithm should strive to find an optimum path for packet transmission within a specified time so as to satisfy the Quality of Service [3]-[5].

For Shortest path problem there are number of search algorithms such as Bellman-Ford algorithm, the Dijkstra algorithm etc. to name a few [2]. Many of these algorithms find the shortest path considering hop count/minimum cost measures. They exhibit unacceptably high computational complexity for communication involving rapidly changing network topologies and involvement of other routing metrics like path reliability, delay, bandwidth etc. for computation of optimal path from source to destination [4], [5]. The selection of routes in large-scale computer communication networks is extremely complex network optimization problems. Such problems belong to the class of nonlinear combinatorial optimization problems most of which are NP-hard.

The problem can be formulated as finding a minimal cost path with greater path fitness that contains the designated destination and source nodes. The Shortest and reliable Path routing problem is a Classical combinatorial optimization problem. Evolutionary algorithms such as, genetic algorithms, neural networks, ant colony optimization etc. promise the solution for such problems, as these have been used successful in many practical applications. Neural Networks and Genetic Algorithms may also not be promising candidates for supporting real-time applications in networks because they involve a large number of iterations in general but will be able to provide the adequate solution. In literature GA is the most popular technique to solve complex multi objective optimization problems.

Researchers have applied GAs to the Shortest Path routing problem, multi casting routing problem, ATM bandwidth allocation problem, capacity and flow

assignment (CFA) problem, and the dynamic routing problem [6]-[13]. It is noted that all these problems can be formulated as some sort of a combinatorial optimization problem. Genetic algorithms are multiple, iterative, stochastic, general purpose searching algorithms based on natural evolution [14], [15]. They aren't instantaneous, or even close, but they do an excellent job of searching through a large and complex search space [16]. The main objective of the current work is to find the best reliable path using personalized GA with various crossover operators for data transmission.

The structure of the paper is organized as follows: A method of representing the solutions is given in section 2. The proposed GA based routing technique is given in section 3. The brief review of crossover techniques is given in section 4. The experimental results of proposed work are presented in section 5. Conclusions are given in section 6.

## 2. Genetic Algorithm Chromosome Representation

To apply GA for any optimization problem, one has to think a way for encoding solutions as feasible chromosomes so that the crossovers of feasible chromosomes result in feasible chromosomes [17]. The techniques for encoding solutions vary by problem and, involve a certain amount of art. Proposed work has considered permutation encoding for chromosome representation. Each gene of a chromosome takes a label of node such that no node can appear twice in the same chromosome. For example, let  $\{1, 2, 8, 10, 12, 14\}$  be the labels of nodes in a 6 node instance, then a route  $\{1 \rightarrow 2 \rightarrow 8 \rightarrow 10 \rightarrow 12 \rightarrow 14\}$  may be represented as (1, 2, 8, 10, 12, 14).

## 3. Routing Technique based on Reliability and Delay Measures

One of the most common problems encountered in networks is obtaining the shortest, reliable, optimum path problem. The objective of the proposed technique is to find the best reliable path in a communication network using personalized GA with historical reliability factor, node success/failure and delay measures. Let us consider a point-to-point communication network modeled by the simple connected graph  $G = (V, E)$ , where 'V' is the set of nodes (or processors or routers) and 'E' is a set of edges (or bidirectional communicational links). Each element  $(u, v)$  in 'E' is an edge joining node u to node v. A path in a graph from a source node 's' to a

destination node 'd' is a sequence of nodes  $(V_0, V_1, V_2, \dots, V_k)$  where  $s = V_0$  and  $d = V_k$ . The proposed technique uses the obtained historical reliability factor generated randomly between 1 and 100, binary representation ("1" for successful transmission of packets and "0" for failure) of node success/failure details in the past 'n' time period and a delay component  $\omega$  at each node in path for the choosing the best reliable path to sent the packets from source to destination. For current work 'n' is taken as 15.

### 3.1 Reliable and Optimized Route Selection

The purpose of the proposed technique is to provide the optimum, reliable route between the source and destination considering historical reliability factor, node success failure and delay constraints.

#### Proposed Algorithm:

1. Perform Initialization of initial Chromosomes/Routes
2. Repeat ( until terminated)
3. Fitness Evolution of each valid chromosome on basis of fitness function
4. Perform selection from evaluated chromosomes using rank selection for crossover
5. Perform crossover using 1-point/2-point/Adaptive techniques
6. Perform mutation with probability 0.1
7. Quit the process if termination criteria (No. of Generations) meet; It can be Time, Minimum fitness threshold satisfied
8. Go to step 2 if Not Terminating

Initially all the possible, connected paths from source to destination, which are the chromosomes of GA, are generated, subsequently from the generated chromosomes, the 'x' number of chromosomes are selected randomly to create the initial population and then the fitness of each chromosomes in the population are calculated. According to the ranking selection with the fitness value, the best 'y' chromosomes are selected for crossover. During crossover operation, we will use the 1-point, 2-point and adaptive crossover to analyze the path fitness. After performing crossover, the repetition of similar chromosomes in the produced offspring as well as repetition of nodes in a chromosome is checked out and duplication is removed. Next, all process from the fitness evaluation to repetition checking are carried out 'C' no of times to obtain the best reliable path.

### 3.1.1 Initial Population Generation and Chromosome Representation

The route table which contains the possible paths from “s” to “d” is generated. Let 'RT' be the generated route table consisting of the possible path from source to destination. The each path in the route table becomes as the chromosome of GA. The gene represents the node while the chromosome represents the network path. Population is the collection of 'x' possible paths selected randomly from the 'RT'. The proposed technique uses the permutation encoding in which each gene represents the node number in a path. Chromosome representation of the possible network path may be as 1-6-7-10-14 which is constituted by nodes and the first node is the source node and the last node is the destination. Hop count of the path will be  $hop = (l - 1)$  where  $l$  the total no of nodes in the path. The fitness function is used to numerically evaluate the quality of the each chromosome within the population.

### 3.1.2 Fitness Evaluation

The genetic algorithm searches for the optimum route with highest fitness where the fitness function is used to assess the quality of a given schedule within the population [18]. The fitness function that involves computational efficiency and accuracy is defined in equation (1):

$$f_i = \frac{\sum_{i=1}^{hop} \mu_i(N) \times \alpha_i(N) \times \frac{\beta_i(N)}{b} \times \sum_{i=1}^{hop} \omega_i(N)}{hop} \quad (1)$$

Where,  $\mu_i(N) = \frac{count(A[N][j] > \lambda)}{n}$  is the historical satisfied reliability ratio of the node 'N' and  $\lambda$  is the minimum required reliability for transforming

the packets [18].  $\alpha_i(N) = \frac{count(B[N][j] = 1)}{n}$  is the historical packet transmission success ratio of the node 'N',  $\beta_i(N)$  is the capacity of the node 'N', 'b' is the packet size and  $\omega_i(N)$  is the delay time of the node 'N'. The fitness of every chromosome in the population is evaluated and based on ranking selection method the best 'y' chromosomes are chosen for crossover operation.

### 3.1.3 Crossover Operations

Three crossover methods 1-point, 2-point and adaptive crossovers are explained below:

**1-Point Crossover:** In this crossover operation choose the parent chromosomes from the population such that, at

least one node is common to both the parent chromosomes [14]. If there is more than one common node, the first occurring common node from left is considered for operation. For example, in Fig.1 (a) the chromosome 'A' and 'B' are the parent chromosomes having the node '7' as common node.



Fig. 1 (a) 1-Point Crossover operator

**2-Point Crossover:** Select the parents from the population on basis of their ranking. Randomly two points are chosen in parent chromosomes [14], [19]. The nodes in between the chosen points are exchanged to generate the children chromosomes.



Fig. 1 (b) 2-Point Crossover operator

**Adaptive Crossover:** From the selected parents A and B, find all the same nodes except source node and destination node, and establish common nodes set  $\Psi$  [20]. If  $\Psi = \Phi$ , there is no crossover operation between A and B. Otherwise, select common node close to source node as a crossover point (for  $\Psi > 1$ , building block hypothesis) from  $\Psi$ , then crossover the part after A and B. For example, if source node is 1 and destination node is 14. Crossover operation is as follows:

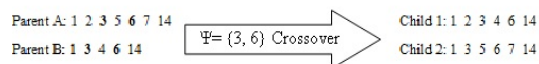


Fig. 1 (c) Adaptive crossover operation

The proposed work uses the 1-point/2-point/adaptive crossover. After performing crossover operation, all the obtained offspring are subjected to repetition checking and duplicate chromosome are removed subsequently and each chromosome is checked for node repetition and maintains the uniqueness of node in a chromosome. Finally the crossover operation provides the best 'z' chromosomes.

### 3.1.4 Selection of the Best Route

The processes in 3.1.2 to 3.1.4 are repeated ‘C’ number of time period with the next population set ‘ $\gamma$ ’ as follows:  $\gamma = z + x'$ , where ‘z’ is the obtained set of offspring from crossover operation and  $x'$  is the next set of chromosomes selected randomly from the route table. After completing the full iterations the best chromosome having the highest fitness is selected from the obtained group of chromosomes. Since the reliability of the nodes having the dynamic nature, the reliability factor of the each node get changes in each time period as follows:  $\hat{A} = \hat{A} + \hat{A} \cdot \xi$ , where ‘ $\xi$ ’ is the reliability deviation factor. The obtained best chromosome represents the best reliable path from the source ‘s’ to the destination ‘s’.

## 4. Related Work on Crossover Operators

Traditionally, GAs have used one-point or two-point crossover [15]. Researchers have also carried out experiments with multi-point crossover: n-point crossover and uniform crossover [21], [22]. With the n-point crossover, n cut points are randomly chosen within the strings and the n-1 segments between the n cut points of the two parents are exchanged. Uniform crossover is the generalization of n-point crossover.

## 5. Experimental Results

This section details the experimentation and performance evaluation of the proposed work. The proposed work is implemented in the MATLAB platform (version 7.12) on Windows 7 platform with Intel Core 2 Duo processor (1.83 GHz) and 2 GB RAM. To illustrate the proposed work consider the example communication network with 15 nodes shown in Fig.2. The maintenance management system of the distribution network maintains the reliability statistics, delay as well as the node failure status statistics for the analysis purpose. The proposed technique uses assumed historical reliability factor, assumed node success/failure status and randomly generated delay.

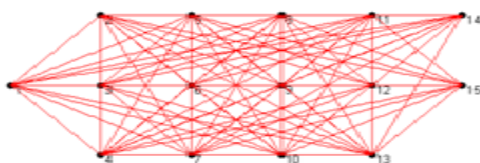


Fig. 2 Communication Network

Let us consider the source node as ‘1’ and the destination node as ‘14’. To find the optimum reliable path, the route table containing the possible ‘p’ paths is generated using a standard route generation algorithm. From that, selected ‘x’ paths forms population subjected to the personalized genetic algorithm, consequently the fitness of the chromosome are calculated, subjected to the genetic operator crossover to produce best offspring and hence produce new population. The above process is repeated for ‘C’ times and finally the best reliable path with high fitness is obtained. The dynamic property of the reliability factor is carried out in each iteration (i.e.) the reliability factor is updated in each time period. The table-1 illustrates some of the possible paths from the source ‘1’ to the destination ‘14’ according to the communication network in the Fig. 2, which are generated using a standard route generation method and table-2 shows the change in path fitness during generations.

Table 1: Routes

No	Path
1	[1,2,3,4,14]
2	[1,2,3,5,14]
3	[1,2,3,6,14]
4	[1,2,3,7,14]
5	[1,2,3,8,14]
6	[1,2,3,10,14]
7	[1,2,3,11,14]
8	[1,2,3,12,14]
9	[1,2,3,13,14]
10	[1,2,3,15,14]

Table 2: Fitness values

Iteration No	Fitness
1	2.2433
2	2.4355
3	2.5217
4	2.6434
5	2.6955
6	2.7123
7	2.7290
8	2.7635
9	2.7994
10	2.8177

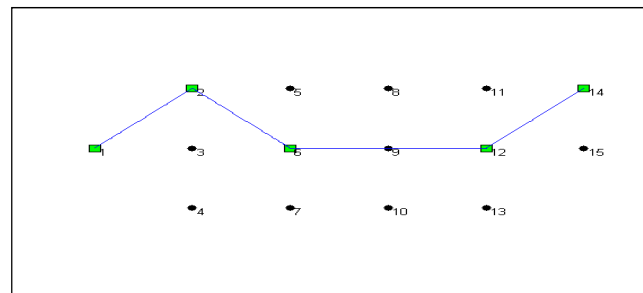


Fig. 3 Reliable and Optimum path

### 5.1 Performance Analysis

The performance of the proposed technique is evaluated in the static environment with the same source node ‘1’ and destination node ‘14’ and its fitness is analyzed using different crossover methods with the Dijkstra algorithm. Comparison is made using delay component in fitness function as well as without delay component in

fitness function with different crossover methods. The table-3 summarizes the fitness values of the various best paths obtained for proposed technique without delay with different crossover methods and Dijkstra algorithm in static environment for different number of iterations. The table reflects that the fitness of the best path obtained using proposed technique is higher than the Dijkstra algorithm. The proposed technique is evaluated for dynamic environment by determining the path with dynamically varying reliability factors.

The proposed technique is evaluated with the different no of nodes and generations. From table-3 it can be observed that adaptive crossover gives better result in path fitness. Overall also from route fitness basis proposed technique represents high performance in the static environment with the Dijkstra algorithm.

Table 3: without delay component in fitness function

30 Iterations			
Fitness	Fitness	Fitness	Fitness
Dijkstra	1-Point Crossover	2-Point Crossover	Adaptive Crossover
1.286	2.5219	2.4268	2.7439
2.2535	2.2535	2.4407	2.8154
1.4521	2.1672	2.3521	3.0407
<b>1.6639</b> <b>(Average)</b>	<b>2.3142</b> <b>(Average)</b>	<b>2.4065</b> <b>(Average)</b>	<b>2.8667</b> <b>(Average)</b>

From table-4 and figure-4 it can be observed that adaptive crossover gives better result in path fitness with delay.

Table 4: with delay component in fitness function

30 Iterations			
Fitness	Fitness	Fitness	Fitness
(Dijkstra)	1-Point Crossover	2-Point Crossover	Adaptive Crossover
1.3214	2.6322	2.7221	2.9774
1.6722	2.4565	2.4903	3.1234
1.5523	2.3345	2.2341	2.6852
<b>1.5153</b> <b>(Average)</b>	<b>2.4744</b> <b>(Average)</b>	<b>2.4822</b> <b>(Average)</b>	<b>2.9287</b> <b>(Average)</b>

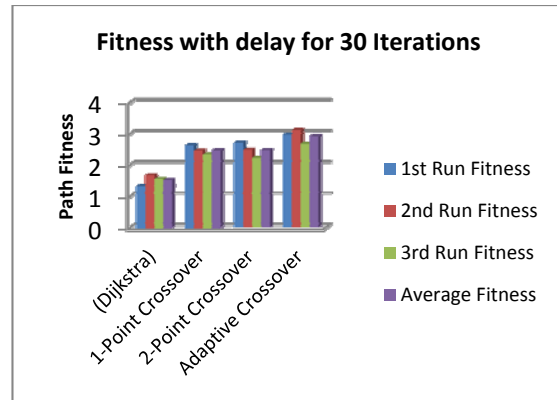


Fig. 4 Performance Comparisons

Convergence Graph for fitness in various iterations using crossover is shown in figure-5 (a) and figure-5 (b) below:

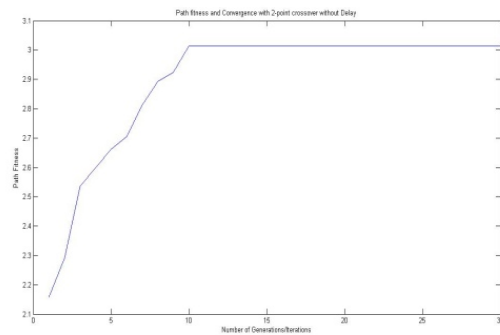


Fig. 5 (a) Convergence Graph

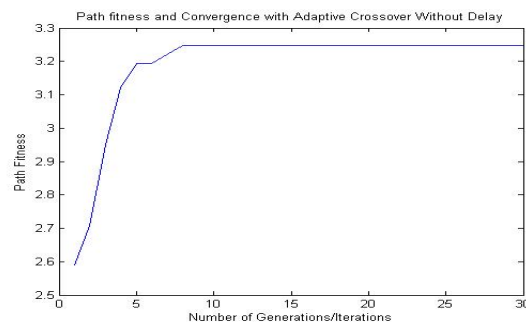


Fig. 5 (b) Convergence Graph

## 6. Conclusion

In this research work, the use of Genetic Algorithm has been proposed to find out the optimal route in packet switched network. The three different types of crossover operators have been proposed i.e. 1-pont crossover, 2-point crossover and adaptive crossover and the performance have been evaluated using a fitness function

that employed the parameters historical reliability, node success/failure and delay at each node. While doing the comparison of performance it has been observed that adaptive crossover has outperformed the 1-point crossover and 2-point crossover. Moreover all the variant of Genetic algorithm i.e. 1-point, 2-point and adaptive crossover have outperformed the Dijkstra algorithm. The researchers are of the opinion that further improvement can be obtained by hybridizing the genetic algorithm with other intelligent optimization techniques such as tabu search and ant colony optimization.

## References

- [1] W.A. Chang, and R.S. Ramakrishna, "A Genetic Algorithm for Shortest Path Routing Problem and the Sizing of Populations," IEEE transactions on evolutionary computation, vol. 6, no. 6, pp. 566-579, Dec. 2002.
- [2] W. Stallng, "High-Speed Networks TCP/IP and ATM Design Principles," Prentice-Hall, 1998.
- [3] M.K. Ali and F. Kamoun, "Neural networks for shortest path computation and routing in computer networks," IEEE Trans. Neural Networks, vol.4, pp.941-954, Nov.1993.
- [4] D.C. Park, and S.E Choi, "A neural network based multi-destination routing algorithm for communication network," in Proc. Joint Conf. Neural Networks, pp.1673-1678, 1998.
- [5] C.W Ahn, R.S Ramakrishna, C.G Kang, and I.C Choi, "Shortest Path routing algorithm using Hopfield neural network," Electron. Lett., vol. 37, no.19, pp.1176-1178, Sept.2001.
- [6] M. Munemoto, Y. Takai, and Y. Sato, "A migration scheme for the genetic adaptive routing algorithm," in Proc. IEEE Int. Conf. Systems, Man, and Cybernetics, pp. 2774-2779, 1998.
- [7] J. Inagaki, M. Haseyama, and H. Kitajima, "A genetic algorithm for determining multiple routes and its applications," In Proc. IEEE Int. Symp. Circuits and Systems, pp.137-140, 1999,
- [8] Y. Leung, G. Li, and Z.B. Xu, "A genetic algorithm for the multiple destination routing problems," IEEE Trans. Evol. Comput., vol. 2, pp. 150-161, Nov.1998.
- [9] Z. Xiawei, C. Changjia, and Z. Gang, "A genetic algorithm for multi casting routing problem," in Proc. Int. Conf. Communication Technology (WCC-ICCT2000), pp. 1248-1253, 2000.
- [10] Q. Zhang, and Y.W. Leung, "An orthogonal genetic algorithm for multimedia multicast routing," IEEE Trans. Evol. Comput., vol. 3, pp.53-62, Apr.1999.
- [11] H. Pan, and I.Y. Wang, "The bandwidth allocation of ATM through genetic algorithm" in Proc. IEEE GLOBE COM' 91, pp. 125-129, 1991.
- [12] M.E. Mostafa, and S.M.A. Eid, , "A genetic algorithm for joint optimization of capacity and flow assignment in packet switched networks," in Proc. 17th National Radio Science Conf., pp. C5-1-C5-6, 2000.
- [13] N. Shimamoto, A. Hiramatsu, and K. Yamasaki, "A dynamic routing control based on a genetic algorithm," in Proc. IEEE Int. Conf. Neural Networks, pp. 1123-1128, 1993.
- [14] D.E. Goldberg, "Genetic Algorithms in Search, Optimization, and Machine Learning" Addison-Wesley, New York, 1989.
- [15] J. Holland, "Adaptation in Natural and Artificial Systems" The University of Michigan Press, Ann Arbor, 1975.
- [16] A.M. Saad, M. Mustafa, N. Amer, and AbuAli "Hybrid dynamic routing protocol for finding an optimal path to routers" International Journal of Academic Research, Vol. 3. No. 2, Part IV, Mar., 2011.
- [17] A.H. Zakir "Genetic Algorithm for the Traveling Salesman Problem using Sequential Constructive Crossover Operator" International Journal of Biometrics & Bioinformatics (IJBB), Volume (3): Issue (6), 2004
- [18] Rakesh Kumar and Mahesh Kumar, "A novel routing technique for computer networks using personalized GA, (Under review process in Journal)
- [19] K.A. DeJong, "Analysis of the Behavior of a Class of Genetic Adaptive Systems" PhD Thesis, University of Michigan, Ann Arbor, 1975.
- [20] W. Xilo-Yan, and L. Yang, "Routing optimization algorithm of mobile Ad-Hoc networks based on Genetic Algorithm" American Journal of Engineering and Technology Research, Vol. 11, No.9, 2011.
- [21] L.J. Eshelman, R.A. Caruana, and J.D. Schaefer, "Biases in the crossover landscape" In Proc. of the 3rd Int. Conf. on Genetic Algorithms, 10-19. Morgan Kaufmann, 1989.
- [22] G. Syswerda, "Uniform crossover in genetic algorithms," In Proc. of the 3rd Int. Conf. on Genetic Algorithms, 2-9, Morgan Kaufmann, 1989.



Rakesh Kumar obtained his B.Sc. Degree, Master's degree – Gold Medalist (Master of Computer Applications) and PhD (Computer Science & Applications) from Kurukshetra University, Kurukshetra. Currently, he is Associate Professor in the Department of Computer Science and Applications, Kurukshetra University, Kurukshetra, Haryana, India. His research

interests are in Genetic Algorithm, Software Testing, Artificial Intelligence, and Networking. He is a senior member of International Association of Computer Science and Information Technology (IACSIT).



Mahesh Kumar obtained his B.Sc. Degree, Master's degree in Science (IT) and Master's degree in Engineering (Computer Science & Engineering) from Kurukshetra University, Kurukshetra. Currently, he is a research scholar in the Department of Computer Science and Applications, Kurukshetra University, Kurukshetra, Haryana, India. His research interests are in Computer Networks, Database systems and Genetic Algorithms.

# Serious Security Weakness in RSA Cryptosystem

Majid Bakhtiari<sup>1</sup> Mohd Aizaini Maarof<sup>2</sup>

<sup>1</sup> Department of Computer Science & Information Systems, University Technology Malaysia, Skudai Johor Bahru, 81310, MALAYSIA

<sup>2</sup> Department of Computer Science & Information Systems, University Technology Malaysia, Skudai Johor Bahru, 81310 MALAYSIA,

## Abstract

Nowadays, RSA is the well-known cryptosystem which supports most of electronic commercial communications. RSA is working on the base of multiplication of two prime numbers. Currently different kinds of attacks have identified against RSA by cryptanalysis. This paper has shown that regardless to the size of secret key and public key, it is possible to decrypt one cipher text by different secret keys RSA algorithm and in excellent condition, there are two similar key at least available in domain of two prime numbers multiplication.

**Keywords:** RSA, Similar Key, Different Secret Key, Encryption, cryptanalysis.

## 1. Introduction

RSA algorithm has invented by Ron Rivest, Adi Shamir and Leonard Adleman (RSA) in 1977 [1]. RSA has categorized in asymmetric key classification and it has capability to supports encryption and digital signature [2].

Currently, RSA is used in security protocols [3] such as:

- TLS/SSL - transport data security (web)
- PGP - email security
- IPSEC/IKE - IP data security
- SILC - conferencing service security
- SSH - terminal connection security

Nevertheless, the RSA is a famous public key algorithm used in the world.

RSA is working on the base of multiplication of two prime numbers. Therefore, number factorization is a serious threatening against RSA. Today, the large numbers factorization is major problem in the world. However, there are a lot of inefficient algorithms available today which will correctly factor big numbers. The idea of RSA can be best depicted in Figure 1.

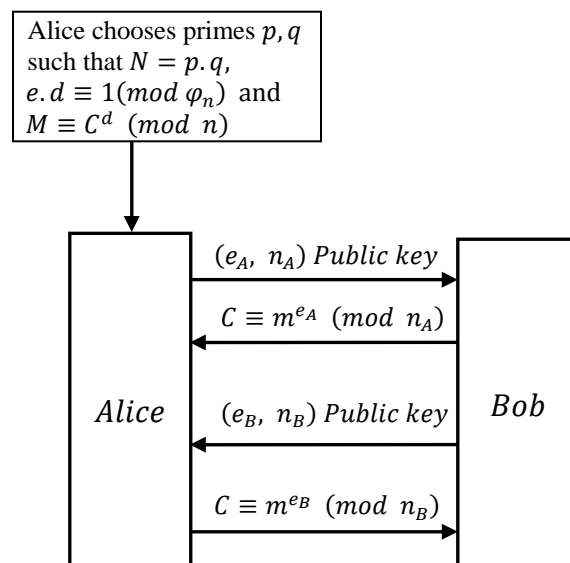


Fig. 1: Diagram of RSA Encryption and Decryption

With consider that RSA cryptosystem works on a given modulus " $n$ " that is the product of two prime random numbers " $p$ " and " $q$ ", a public exponent " $e$ ", and an element  $C \in Z_n$ , users find " $m$ " such that  $C = m^e \pmod{n}$  and a private exponent " $d$ " that should satisfy  $m = C^d \pmod{n}$ . Therefore, at the first view the following items should be considered:

- The bit-size of " $n$ "
- The size of the public exponent " $e$ "
- The size of the private exponent " $d$ "
- The factorization of " $n$ "
- The features of two big random prime numbers

In 1977 Ron Rivest said that factoring a 125-digit number would take 40 quadrillion years. In 1994 RSA129 was factored using about 5000 MIPS-years of effort from idle CPU cycles on computers across the Internet for eight



months. Today, it is possible to factorize 192 digit numbers easily [4].

However, there are some algorithms namely the Trial division, Pollard's Rho algorithm, Pollard's p-1 algorithm, Williams' p+1 algorithm, Lenstra elliptic curve factorization, Fermat's factorization method and Special number field Sieve that can factorize big numbers into prime numbers. But this paper concentrates on the other viewpoint in RSA which explore one of the big weaknesses of RSA cryptosystem.

This paper has shown that regardless to the size of secret key and public key, it is possible to decrypt one cipher text in RSA algorithm by different secret keys.

## 2. Background

There are many kinds of attacks have known against RSA algorithm. The most well-known of them are listed as follows:

- Common modulus
- Blinding
- Small encryption exponent “e”
- Small decryption exponent “d”
- Forward search attack
- Timing attack
- Multiplicative properties
- Cycling attack
- Message concealing
- Faulty encryption attack
- Factoring the public key

It should be notice that no attack algorithm can break RSA cryptosystem in efficient manner. Most attacks appear to be the result of misuse of the system or bad choice of parameters. Analysis of the known attacks shows that RSA has not been proven to be unbreakable, but having survived a great deal of cryptanalytic security over the last thirty years [5].

## 3. New Security Weakness in RSA

According to Fermat's little Theorem on the probable prime number which stated if  $p$  is a prime and  $a$  is an integer coprime to  $p$ , then  $a^{p-1} - 1$  will be evenly divisible by  $p$ . Therefore, in the notation of modular arithmetic:

$a^{p-1} \equiv 1 \pmod p$ . Otherwise,  $p$  is composite number.

With consider that RSA algorithm is working on the base of two prime numbers  $p, q$ . The Fermat's theory can be expanded in some part of RSA algorithm as follows:

$$n = p \cdot q$$

$$\varphi = (p - 1)(q - 1)$$

$$2^{p-1} \pmod p = 1 \tag{1}$$

$$2^{q-1} \pmod q = 1 \tag{2}$$

From Eq. 1 and Eq. 2:

$$2^\varphi \pmod n = 1 \tag{3}$$

On the other hand, with consider that  $(p - 1)$  and  $(q - 1)$  are dividable to 2, therefore, in second step:

$$2^{\frac{\varphi}{2}} \pmod n = 1 \tag{4}$$

Eq. 4 is proving that, at least it is two same fields available between 0 to  $n$ . In other word

$$2^{\varphi+\Delta x} \pmod n = 2^{\frac{\varphi}{2}+\Delta x} \pmod n \tag{5}$$

Eq. 5 is proving that at least two same fields are available in domain of “ $n$ ”. It means that any number between 1 to  $\frac{\varphi}{2}$  have same properties to any number between  $\frac{\varphi}{2}$  to  $\varphi$ . With consider that “ $e$ ”, “ $n$ ” as public key and “ $d$ ”, “ $n$ ” as secret key have located in domain of  $\frac{\varphi}{2}$ .

Therefore, as it has shown in Figure 2, there are at least two “ $d$ ” and two “ $e$ ” are available which have exactly same properties concerning to RSA cryptography operations.

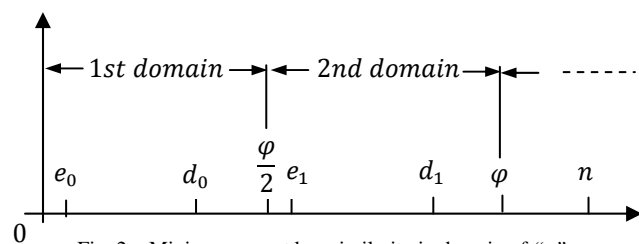


Fig. 2: Minimum secret key similarity in domain of “ $n$ ”

$$C = m^{e_0} \pmod n = m^{e_1} \pmod n \tag{6}$$

$$m = C^{d_0} \pmod n = C^{d_1} \pmod n \tag{7}$$

As it has shown in Eq. 6 and Eq. 7, it is possible to decrypt one cipher text by two separate secret keys. It should be noticed that the Eq. 6 and Eq. 7 are valid if and only if that two prime numbers  $(p, q)$  have unique specification. Otherwise, the numbers of secret keys which can decrypt one cipher text are more than two in domain of “ $n$ ”. The

unique specification of two prime numbers have identified as follows:

- $\frac{p-1}{2}$  must be prime number.
- $\frac{q-1}{2}$  must be prime number.

Those numbers which have passed above condition successfully called “strong prime” in this paper. If and only if,  $p, q$  cannot pass the strong prime conditions successfully, the number of similar secret keys (which are more than two keys) are depend to combination of factorization items of  $p$  and  $q$ . The following example shows one sample of this procedure, if  $p$  and  $q$  cannot pass above condition successfully.

$$p = 401 ; q = 281 ; n = 112681 ; e = 3$$

$$p - 1 = 2^4 * 5^2$$

$$q - 1 = 2^3 * 5 * 7$$

With consider that, public key and secret key should be generated on the base of  $\phi$ . Therefore, selected prime numbers ( $p, q$ ) have 40 secret key that each one can decrypt all of cipher texts that encrypted by one public key. It is because  $\frac{p-1}{2}$  and  $\frac{q-1}{2}$  are not prime number. In this example, Table 1 shows some of same secret keys for a public key (3, 112681).

Table 1: One public key with different similar secret keys ( $n=112681$ )

Public Key	Secret key
$e = 3$	1867, 4667, 27067, ...

It is important to notice that finding  $p$  and  $q$  in such a way that  $p \pm 1$  and  $q \pm 1$  to have large prime factor are not enough conditions that RSA laboratories advised [6-8]. It is because; mentioned conditions cannot solve similarity keys in domain of “ $n$ ”. They tried just to reduce this serious weakness in RSA.

On the other hand, it should be mention that the numbers of strong prime numbers are very limited. According to study that generated Figure 3, number of strong prime numbers for more than 256 bits (77 digits) are very limited. Therefore, most of pair prime numbers cannot provide the defined condition. Also this why that RSA laboratory advise to select large prime numbers with large prime factor [7].

Figure 3 shows the density of strong prime numbers up to 20 digits (64-bit). As it has shown, they are very limited and it is possible to generate those numbers and store them in one data base centre for efficient attack against RSA.

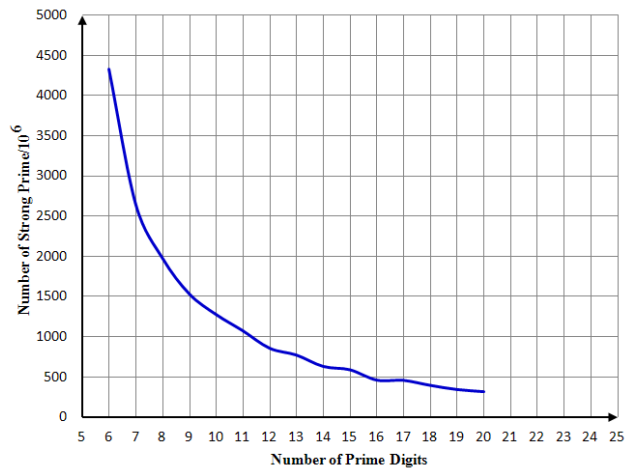


Fig. 3: The density of strong prime numbers for RSA

### 3.1 Number of similar secret key

By extending the properties of two prime numbers multiplication which RSA cryptosystem is following, there are many similar secret keys are available out of domain of “ $n$ ”. It means that even by selecting strong prime numbers, there are infinite secret keys exist which located in  $\mathbb{R}$ . The distance of each secret key that has shown in Figure 4 is equal to  $LCM((p - 1), (q - 1))$ .

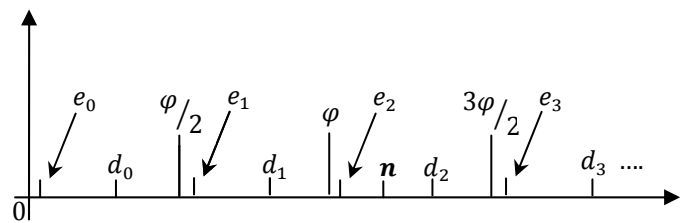


Fig 4: Number of similar secret keys in Real Number field

Figure 4 shows that there are infinite “ $d$ ” are available which can decrypt a message that encrypted by  $e_i ; i > 0$ . It means that for  $C_1 = m^{e_i} \text{ mod } n ; i > 0$ , it is possible to find plain message by  $m = C_1^{d_j} \text{ mod } n ; j > 0$ . The minimum distance between two secret key is equal to  $d_{i+1} - d_i = LCM((p - 1), (q - 1)) ; i \geq 0$ .

### 3.2 New threatening on RSA

According to discussion in Section 3 by Eq. 5, due to structure of RSA which has based on multiplication of two prime numbers, there is new vulnerability available in RSA; even two prime numbers are strong prime. As it has shown in following equation:

$$p = k.q \Rightarrow n = k.q^2 \quad (8)$$

$$\sqrt{n} = q\sqrt{k} \Rightarrow \begin{cases} q\sqrt{k} < p \\ q\sqrt{k} > q \\ q\sqrt{k} < \frac{p+q}{2} \end{cases} \quad (9)$$

$$\text{From Eq. 9} \Rightarrow \frac{p+q}{2} - \Delta x = q\sqrt{k} \quad (10)$$

Therefore,

$$2^{(n-q\sqrt{k})} \bmod n = 2^{\left(\frac{p+q}{2} + \Delta x\right)} \bmod n \quad (11)$$

In this step, it is important to find one feature from point of  $\frac{p+q}{2}$ . According to Eq. 3 and Eq. 4, we can write Eq. 12 as follow:

$$2^{\frac{n+1}{2}} \bmod n = 2^{\frac{p+q}{2}} \bmod n \quad (12)$$

With consider that, the left side of both Eq. 12 and Eq. 11 are determined and right side of Eq. 10 is determined too. Therefore, it is concluded that the maximum security level for RSA is equal to find  $\Delta x$ . while  $\Delta x$  is very smaller than amount of “ $q$ ”. However, there are different methods are available to find  $\Delta x$ .

Figure 5 shows a sample flow diagram to finding  $\Delta x$ . In method which has shown in Figure5, basic instruction are just shift left ( $x = x * 2$ ), take modulo  $n$  and simple addition. However, there are different methods exist to find  $LCM(p - 1, q - 1)$  by Eq. 12.

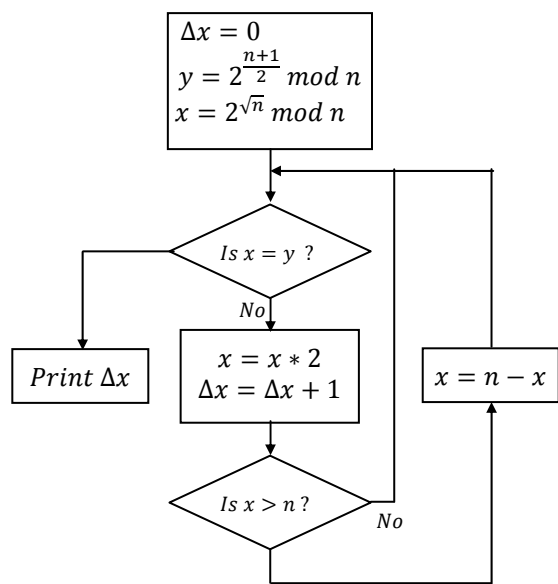


Fig. 5: One sample flow diagram to finding  $\Delta x$

## 4. Conclusion and Future work

This paper proved that RSA cryptosystem has at least two similar secret key in domain of “ $n$ ” for all of cipher texts and infinite similar secret key are exist out of domain of “ $n$ ”. Also this paper proved that the maximum security level of RSA is not equal to bit-length of “ $n$ ” and for any length-bit of “ $p$ ” and “ $q$ ”. According to study of this paper, the security level of RSA cryptosystem is smaller from digit length in comparison to each of two selected prime numbers.

Currently, it is not correct evaluation between different cryptosystem and RSA. Finding an efficient method to obtain  $\Delta x$  or  $\varphi_n$  by Eq. 12, and then evaluating RSA cryptosystem from security level view point can be good future work to evaluate RSA cryptosystem.

## References

- [1] Rivest, R.L., A. Shamir, and L. Adleman, *A method for obtaining digital signatures and public-key cryptosystems*. Communications of the ACM, 1978. **21**(2): p. 120-126.
- [2] ElGamal, T. *A public key cryptosystem and a signature scheme based on discrete logarithms*. 1985: Springer.
- [3] Sonsare, P.M. and S. Sapkal. *Stegano-CryptoSystem for Enhancing Biometric-Feature Security with RSA*. 2011.
- [4] Cavallar, S., et al. *Factorization of a 512-bit RSA modulus*. 2000: Springer.
- [5] Salah, I.K., A. Darwish, and S. Oqeili, *Mathematical attacks on RSA cryptosystem*. Journal of Computer Science, 2006. **2**(8): p. 665-671.
- [6] Rivest, R. and R.D. Silverman. *Arestrong'primes needed for RSA,*. 1997: Citeseer.
- [7] Silverman, R.D., *Fast generation of random, strong RSA primes*. CryptoBytes, 1997. **3**(1): p. 9-13.
- [8] Gordon, J., *Strong RSA keys*. Electronics Letters Published by IEEE, 1984. **20**(12): p. 514-516.

# Two Automated Mechanisms to Create eLectures and to Videotape Regular Lectures

Nael Hirzallah

Faculty of Information Technology, Applied Science University  
Amman, Jordan

## Abstract

Many institutes are offering lectures that students may watch online. These lectures vary in type from true multimedia sophisticated documents to eLectures to videotaped regular lectures that may include auto generated closed caption. This paper studies some formats that can be easily composed by instructors. Whether adopting the eLectures format or videotaped lecture format, the paper proposes two automated mechanisms to create such lectures without the intervention of a camera or video mixer operator. The first mechanism depends on the absence of mouse motion events; while the second mechanism depends on the instructor's facial and hand detections.

**Keywords:** *Multimedia, eLearning, Open Course Ware, Content Development, Authoring tool*

## 1. Introduction

The major challenge in the implementation of eLearning is the content development process. From one point of view, it is very expensive. From another point, it is time consuming. One may either develop an eLearning platform or choose one from the market to install in the institute or company, but developing the content that is customized to specific audience is another story.

Initially, groups that include developers, designers, material specialists, and testers get together in order to develop an online course. Such a team may have to hold extensive meetings to mainly bridge the communication gap among the various poles of the team. This is because the material specialists are usually from the academic sector, while the others are from the industrial sector. This, by itself, is considered sometimes the first road-block that the project manager must overcome so all can talk the same technical language. The team leader may take sometimes expensive measures to overcome this problem and move forward. However, replacing the material specialist is usually the last option to take. Once the problem is overcome, a complete course would then be split into modules. Once the development and testing cycles are finished and during the integration phase, they will be recombined back. This cycle will be repeated for the second course and so on. However, for a different

course, the team may have to work with another material specialist, and the whole process would then be repeated from bridging the communication gap phase. This was found out to be a very expensive process in terms of money and in terms of time. Therefore, for these reasons, many institutes are ending up implementing an eLearning platform that is rich with features but poor in real content.

The other choice for these institutes is to train their material specialists on using content development tools. Based on the chosen tools and the backing up policies adopted, the decision may end up to be the right one or it would not reach the goal desired by the institute. Many authoring tools are available off-shelves, such as Adobe Captivate [1]. To select the right tool, one should consider the time and efforts the material specialists would need to use it well. It should also consider the time the specialists may need to develop a complete online course. In most cases, the easier to learn, the faster to use the tool is, but the fewer features would be used. Authoring-on-the-fly tools, AOF, [2,3] were developed to satisfy this purpose.

The third and last choice offered to these institutes is to videotape their regular lectures in order to produce online lectures using simple video format. A group of these online lectures will form a complete online course. Three different models are available for videotaping a lecture based on the number and type of cameras used. The different number and types of cameras are: more than one fixed-location camera with a mixing device, a single moving Camera, and a single fixed-location wide lens camera. In the first two models, a video editor/mixer person and a camera-man may be needed.

Due to the high cost of developing sophisticated multimedia lectures and the speed of the eLearning marathon among educational institutes, many are adopting the AOF and videotaping approaches discussed earlier for online lectures development. For AOF, one of its main motivations when compared with videotaping regular lectures is to minimize the bandwidth and storage capacity needed. In AOF, the output frame is usually composed of

many window regions (similar to Figure 1). These regions include images and hyperlinks along with one or two smaller size videos. It is considered to produce a little more sophisticated outputs than videotaped regular lectures. But, with better new video compression techniques and faster home internet speed, storage capacity and bandwidth do not anymore represent a problem. Thus, more and more universities are adopting the videotaping approach.



Figure 1 Slides and instructor view taken from [4]

This paper studies both approaches, namely the AOF and videotaping by presenting the cons and pros of each. It then proposes an automated simple mixed approach between the AOF, the more sophisticated approach, and that of the videotaping approach, in order to benefit from the cons of both approaches. It also presents the results of a survey performed on a group of students from various fields regarding their perception to the kind of video they believe would help them stay focused during an online lecture. Finally, the paper concludes with the last section.

## 2. Overview

Content development is by far the most challenging task in eLearning. It is a time consuming and costly process. One may categorize the approaches used for this task based on how sophisticated, interactive, and whether it is a true multimedia or not the outcome is. A document is said to be a True Multimedia if it contains some kind of interaction that may change the presenting scenario as well as some animation. Thus, a traditional video cannot be called a true multimedia. True multimedia eLearning documents enjoy many advantages over other less sophisticated documents. Such advantages include being concise, attractive, and the topics covered are usually well presented. Besides, they can be compressed at a much higher ratio than regular videos due to the animations that

they include. However, due the high cost of developing a true multimedia document and the very lengthy process to produce one online course, they have forced many to focus on opportunities using AOF and videotaping approaches.

### 2.1. Videotaping

Videotaping is the easiest way to generate content for any eLearning platform. Traditional videotaping is as simple as putting one or more cameras in a classroom to record the lecture. During a traditional regular lecture, the instructor mainly does nothing extra. The instructor may either ignore the presence of the camera or deal with it as a yet another silent student. In a classroom, the instructor may do one or more of the following at the same time:

1. Reading or Explaining something written on the board or projected on the screen
2. Writing on the board
3. Speaking to students or Listening to their questions

When using one camera, there are three options to use in order to capture all the notes written by the instructor on the board and keep them readable. The first option is to use a small board that would represent the video frame boundaries. The other option is to hire a camera man during the lecture who knows what, how, and when to capture the notes (samples of such an approach can be spotted easily by MIT Open Course Ware system. Figure 2 shows a snapshot from a lecture conducted by Prof. Eric Grimsona and Prof. John Guttag from MIT-OCW).

### 2: Branching, Conditionals, and Iteration



Figure 2: MIT Open Course Ware system

While the last option is to use a camera with motion detection and ask the instructor to keep paying attention to it, so the instructor would stay next to the notes or area of the board being referred to. However, this may require the instructor to move back and forth often and quickly, thus, deviating from the main purpose of catching the notes clearly.

On the other hand, when using more than one camera, there are just two options. The first is to do an A/B mixing during the lecture, while the other option is do it after the lecture. In both cases, you will need some human intervention other than the instructor. Samples of such approach can be found within the NPTEL program [5].

The drawbacks of such an approach are two folds. Firstly, you need a human intervention in most of the scenarios discussed above. In other words, you need a camera man or a person to do the video mixing from the various cameras inputs. Secondly, the length and number of videotaped lectures to cover a complete course are usually large. This is due to the following reasons:

1. In Videotaped lectures, some of the topics scheduled to be covered in a certain lecture may get postponed whenever an attending student asks the instructor to repeat a certain topic.
2. The speed the information being delivered by the instructor at may vary based on the instructor's sense of the way the students perceive the information.
3. Unlike in eLectures, in videotaped regular lectures, the instructor usually takes some time to interact with the students as a kind of Keep-Alive messages, such as throwing questions and hearing answers.

In the case when the instructor uses Softnotes (notes being added or written on slides through a computer) rather than Hardnotes (notes written on a board), AOF becomes a better approach to use to generate online lectures. AOU for instant uses mainly Elluminate Live system [6] (kind of online conferencing system) to deliver and record their online classes.

## 2.2. Authoring on the Fly

Although videotaping a traditional lecture can also be called AOF, but AOF usually refers to a little more sophisticated output which may include more than one region within the output video frame. The simplest output that can be composed from the content of the videotaping approach is to assign one video region to the camera focused on the board, and another region to the camera with a wide shot showing the instructor movements and gesturing. Furthermore, certain positions of the video are then bookmarked and labeled according to the topic being discussed. Consequently, a region that displays the list of the bookmarks will also be included. Examples of such approach are those listed in [4]. Figure 1 shows a sample of such a lecture taken from [4] with the slide show screen used rather than a board.

It is worth to mention that although the output in this example (Figure 1) may look like a complete video played by RealPlayer [7], but in fact, only the video region within the output frame is the real video; while the other regions may be images, scripts, or hyperlinks integrated together using a scripting language, such as SMIL [8]. The main reason for this was to reduce the bandwidth needed. For that, only discrete snapshots of the softnotes or hardnotes, were added. However, we believe that showing discrete or even significant shots of the notes being written by the instructor rather than continuous video of the notes while being added gradually is harder on the students to follow.

Another example of AOF is classroom 2000 [9]. Soft or hard notes written on the white board are being taken, power point slides are being copied, and both the video and audio from the class are being recorded. Such output includes three to four regions within the output video frame.

## 3. Merged Approach

In this paper, we will distinguish between the following three terms of online lectures: Videotaped lectures, AOF lectures, and eLectures. Videotaped lectures represent the output of the in-class cameras recording a complete lecture. AOF lectures, are the output of AOF tools when the lecture is delivered before some students (even if remotely, such as using Elluminate Live). In many cases, it is basically a Videotaped lecture that is integrated with notes generated by some form of video processing or electronic boards. While eLectures are those lectures that are generated by AOF tools when the lecture is delivered *outside* a class room, i.e. with the absence of students. In the next section, more elaboration on eLectures will be presented.

### 3.1. eLectures

eLectures generated outside the lecture rooms are supposed to be more concise than videotaped lectures. Storage and bandwidth needed for a complete course using videotaped lectures are magnitude more in size than if saved as eLectures. At our university for instance, it is found that a twenty minute eLecture is worth up to five hours of videotaped lectures. Also, when comparing with AOF lecture of same duration, although the AOF lectures need much less storage capacity and bandwidth (as mentioned earlier in paragraph 2 of section 2.2), but the same argument about not being concise if compared to eLectures still holds for AOF lectures as well.

For Distance Learning students, videotaped lectures or even AOF lectures may be more beneficent than

eLectures, especially if watched for the first time. For other students who have prior knowledge about the topic, or have watched the video before, it might be boring or time consuming to use them to review the topic or to search through for a particular subtopic.

While the overlap between the videotaping and AOF approaches are high, this paper describes a model that uses bits and pieces from both to generate eLectures. In this model, the instructor speaks before a camera with the absence of students. He/She uses a tablet PC, or a laptop with a pen-like mouse, to replace the white board or the screen. This model is used by some institutes, such as the some lectures from NPTEL [5], in which the video is being switched manually at certain times between a close-up view of the instructor and the laptop screen with the Softnotes being added. Figure 3 shows a snapshot of an eLecture by Dr. Chitralekha Mahanta from IIT. At some other instant, you may see a close-up view of Dr. Mahanta in Full-screen mode. Thus, the video shows either the screen of her laptop or a close up view of the instructor speaking to the camera.

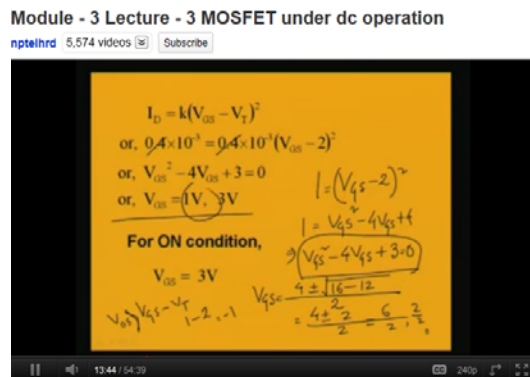


Figure 3: Lecture by Dr. Chitralekha Mahanta from IIT

Using this approach, we may have two different forms based on the location and size of the video window showing the instructor. The forms are as follows:

1. A Full-screen switch between both windows: Close-up view of the instructor and the slide show screen, as in the Figure 3.
2. A small window of the instructor continuously displayed at one of the corners of the screen.

### 3.2. The Instructor in the Video

The most desired components of any output of both approaches mentioned in the previous section are as follows:

1. Clear audio (at least of the instructor)
2. Clear and continuous view of the Notes; whether soft or hard.

3. View of the instructor; either a close up or wide

While the first two components are agreed upon by all as a must for online lectures, you may find a group of people who would debate the third component. They would claim that seeing the instructor adds no value; or more politely, they would claim that the value added by seeing the instructor does not worth the extra storage capacity and bandwidth that they would need. Fortunately, with the cheaper and faster bandwidth and larger storage capacities, it is becoming easier for such a group to be convinced otherwise. Overcoming this obstacle, the issue of a close-up versus wide view of the instructor becomes debatable.

You may find a group of students who would prefer a close up view of the instructor, while others would prefer a wide view. In [4], a survey was performed on a group of students to find their preference towards the existence of such a window. The samples of eLectures that were displayed to them were A, B, and C as follows:

- A. Slides and notes only
- B. Slides and notes with a small window showing a wide view of the instructor at one corner
- C. Slides and notes with a small window showing a close-up view of the instructor at one corner

It was found that the students were better focused with the third sample where a close-up view of the instructor was displayed at all times at one corner of the video window, along with the slides and lecture notes. Figure 4 shows a snapshot of the third sample.



Figure 4 Notes, slides and close up view taken from [4]

Furthermore, we believe that switching the Full-screen mode between the screen and the instructor acts as a Keep-Alive message with the students. Thus, it helps them stay more focused. For that, we have adopted the mixed format between Full-screen switch form and the corner form. However, for this eLecture model, a camera man or a director who would decide when to do a video switch is needed. To overcome this problem, we have developed an automated tool for the instructor to use which will

generate the desired eLecture without the need for another person. The flow chart of the switching algorithm used by the tool is depicted in Figure 5.

As shown in the flow chart of Figure 5, once the system starts or upon a mouse idle time event for at least  $T_i$  (i for idle) seconds, the focus (or full-screen) will be granted to the window showing the instructor. The close-up view of the instructor should last for at least  $T_t$  (t for teacher) seconds. The focus will be granted to the slide-show whenever there is a mouse significant move, or a mouse button being click which indicates that a softnote is about to be written. The focus on the slide-show should last for at least  $T_s$  (s for screen) seconds, and so on. With this simple strategy, it is believed the students will stay focused and the softnotes will not be missed. Note that even when slide-show screen is in Full-screen mode, the close-up view of the instructor is placed in a small window on one corner of the full-screen.

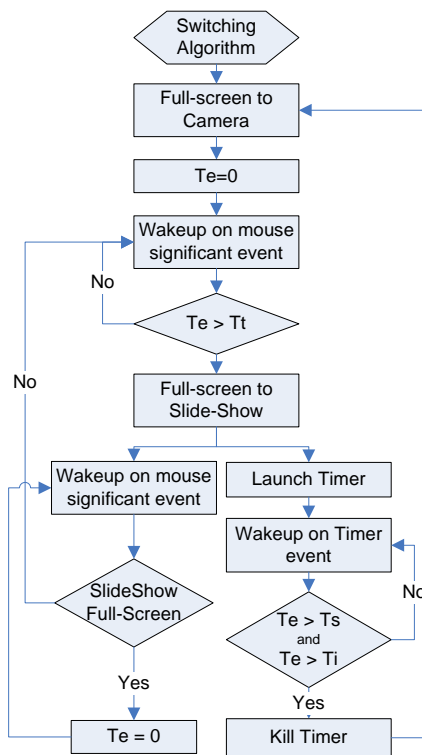


Figure 5 Flow chart of the mixing algorithm

### 3.3 Auto Videotaped Lectures

Many instructors are not used to give lectures with the absence of students. In fact, they prefer to produce long Videotaped lectures rather than composing more concise eLectures. Thus, after all what has been said on eLectures,

videotaped lectures are still considered a fast way to produce lectures online among the vast majority of instructors, as there is no need for much preparation. However, as mentioned earlier, the camera man inside the classroom must know what, how, and when to focus on the instructor or the board. In this section, we propose a way to auto zoom and auto pan the camera in response to the instructor's normal sign language captured by the camera.

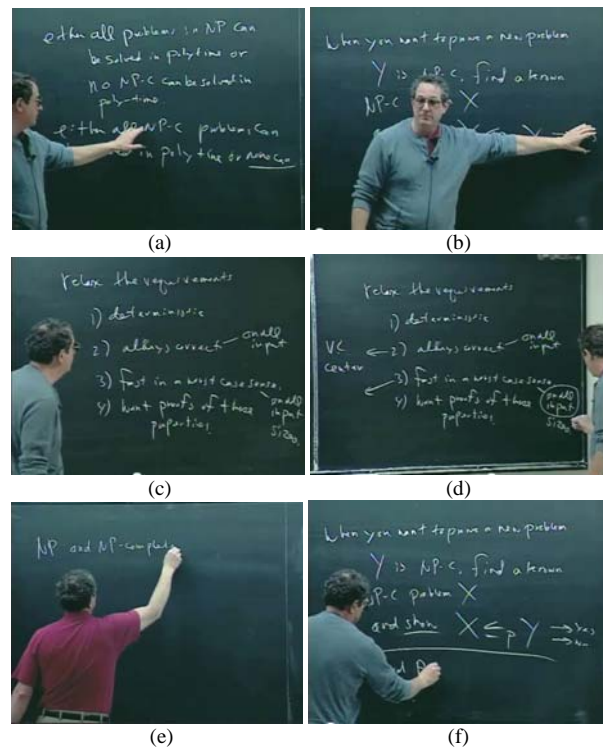


Figure 6 Writing and Explaining

To understand the proposed algorithm, let's first classify the behaviors of the instructor. The Reading, Writing and Speaking behaviors of the instructors mentioned in section 2.1 are illustrated in Figure 6 a to f which shows Prof. Dan Gusfield from UC Davis in some Youtube videos produced by UC Davis [10]. Figures a to c, shows the instructor when Reading from the board in various positions including: sideways, facing audience, and facing the board, respectively. In both figures a and b, the instructor is using his hand to point at the text he is referring to.





Figure 7 Speaking

Figures d to f, show the instructor while writing on the board. Finally, Figure 7a and 7.b show the instructor while speaking to the audience. The major signs to look for in these figures are as follows:

1. An extended hand
2. A face
3. The absence of motion (instructor disappears from the frame)

To detect these from the video, edge and facial detection algorithms are applied. For instance, for edge detection, there are many ways to do so. However, the most may be grouped into two categories: gradient and Laplacian.

The gradient method detects the edges by looking for the maximum and minimum in the first derivative of the image. A pixel location is On or declared an edge location if the value of the gradient exceeds some threshold. Edges will have higher pixel intensity values than those surrounding it. So, once a threshold is set, you can compare the gradient value to the threshold value and detect an edge whenever the threshold is exceeded.

Furthermore, when the first derivative is at a maximum, the second derivative is zero. As a result, another alternative to finding the location of an edge is to locate the zeros in the second derivative. This method is known as the Laplacian. A zero-crossing edge operator was originally proposed by Marr and Hildreth [11]. A method based on the gradient method could be found in [12] while another on the Laplacian method in [11]. Hand detection is based on both edge and motion detection. Hand is simply considered as the non-vertical block that is on high motion. Yet, an overview on existing Facial and Hand detection algorithms is beyond the scope of this paper.

The ideal scenario to record a lecture is to use a wide board that may be split by vertical straight lines into segments of around 1.5 meters wide each, as shown in Figure 9. The camera is mounted on a motor and calibrated based on these straight lines. The maximum zoom is set to cover one complete segment. While the minimum zoom is set to cover the whole board.

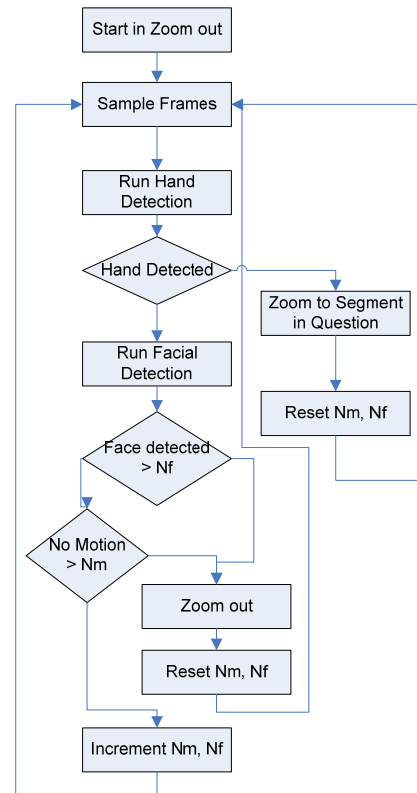


Figure 8 Second Mechanism Flow Chart

The basic camera actions to consider are zooming and panning. To understand the signs that trigger these actions, consider the follow chart depicted in Figure 8.

Frames that are periodically sampled from the video stream are processed for all three signs: Hand detection, Facial Detection and absence of motion. The Zoom out action occurs when either the instructor disappears (absence of motion) for duration  $> N_m$  or the face of the instructor shows up continuously for a duration  $> N_f$ . These parameters refer to the number of sampled frames.

The zoom and panning actions occur when the hand sign appears focusing on the segment of which the hand is pointing at. Thus, with the existence of an extended hand, the facial detection becomes secondary.

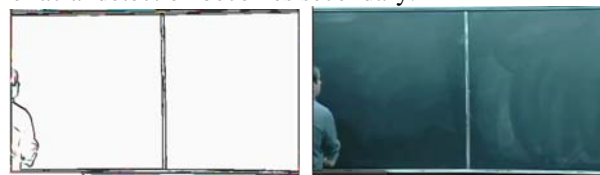


Figure 9 Segmenting the Wide Board by Lines

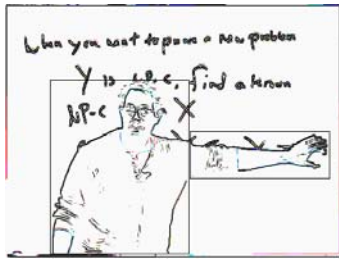


Figure 10 Hand Detection

## 4. Analysis And Survey

Few regular lectures were videotaped and in most cases the camera motor responded to the correct signs. Unlike the Facial and Motion detection algorithms used, the Hand detection was the weakest. Figure 10 shows a successful hand detection illustration to that of Figure 6.b. Cases in which the hand detection algorithm missed were mostly similar to that represented in Figure 6.d. Fortunately, the proposed mechanism does not zoom out or pan easily. Missing hand detection does not affect as much as detecting false ones; and these are rare. Enhancing Hands detection algorithm is left for future work.

A different, yet more focused survey similar to that in [4] was performed. A group of students from different fields, namely Business, and IT, were chosen to select which among the three different models of online lectures would help them to:

1. Stay more focused when watching the lecture for the first time
2. Watch the lecture again for the purpose of reviewing the material or searching through for a specific subtopic.

The three models presented to the students were: a videotaped lecture similar to that of MIT OCW, and second and third models are eLectures similar to that of Dr. Chitralkha Mahanta [5] that switches between the instructor view and the slide-show screen. Both the second and the third differ in the instructor view: the second with a close-up view of the instructor while the third with a wide view. The number of randomly selected students was ten from each field. Each group was given two samples for each model. The first was in the same domain as the students', while the other sample was in a different domain than theirs. The duration of each lecture was about 15 minutes long (the VideoTaped was a little longer). In summary, the results were biased towards eLectures with the close-up view of the instructor.

## 5. Conclusion

Content development process has become the most challenging task in implementing eLearning. The process outcome ranges from being very sophisticated and expensive, such as producing interactive multimedia documents where animation, quizzes, and multiple scenarios exist, to simple videotaped lectures. This paper argues that with the faster home internet speed and the availability of higher bandwidth and storage, videotaped lectures are becoming more popular. However, the drawbacks of these lectures were discussed in this paper. They included their lengths and the need for a camera man and/or a director to do video editing and mixing to produce them. In simple words, a 45 hours course will need about 45 hours of video. Consequently, the paper presented an approach similar to that being used by some institutes to produce eLectures that is more concise in length. A course with 45 contact hours may need less than 4 hours of eLectures. Moreover, the paper proposes a simple way to automate the video mixing process needed for this approach to keep the students more focused, as being claimed by the results of the survey being performed on a group of students. The simplicity is the main feature of this proposal. Therefore, by adopting this approach, more eLectures will be available in less time at a lower price than what most institutes would expect.

## Acknowledgments

The author wishes to acknowledge the financial support received from Applied Science University for publishing this paper.

## References

- [1] Adobe Captivate 5.5 [www.adobe.com/Captivate](http://www.adobe.com/Captivate)
- [2] Ottmann, Th. & Müller, R. (2000). The "Authoring on the Fly"-System for Automated Recording and Replay of (Tele)presentations. ACM/Springer Multimedia Systems Journal, Special Issue on "Multimedia Authoring and Presentation Techniques", Vol. 8, No. 3.
- [3] N Hirzallah, "Developing eLearning Content Considering Various Video Scenarios" The Asian Conference on Education 2009 Osaka Japan, October 24-25 2009.
- [4] N Hirzallah, W Albalawi, A. Kayed, S. Nusair "A simple algorithm to enrich eLectures with instructor notes" ACM Multimedia Tools and Applications: Volume 49, Issue 2 (August 2010), Page 259.
- [5] National Programme on Technology Enhanced Learning <http://nptel.iitm.ac.in/>
- [6] Elluminate Live , blackboard Collaborate™, <http://www.blackboard.com/Platforms/Collaborate/Overview.aspx>
- [7] RealPlayer application, [www.real.com](http://www.real.com)

- [8] Synchronized Multimedia Integration Language (SMIL) 1.0 Specification, W3C Recommendation, 15-June-1998 (<http://www.w3.org/TR/REC-smil/> )
- [9] Gregory D. Abowd: "Classroom 2000 - An Experiment with the Instrumentation of a Living Educational Environment", IBM Systems Journal, Special issue on Pervasive Computing, Volume 38, Number 4, pp. 508-530, October 1999.
- [10] professor Dan Gusfield from UC Davis computer science on Youtube [http://www.youtube.com/watch?v=MZSMcspzj\\_E](http://www.youtube.com/watch?v=MZSMcspzj_E)
- [11] Marr, D., and Hildreth, E. "Theory of Edge Detection," Proceedings of the Royal Society London 207 (1980) 187-217. A.B. Smith, C.D. Jones, and E.F. Roberts, "Article Title", Journal, Publisher, Location, Date, pp. 1-10.
- [12] [12] A. K. Cherri and M. A. Karim, "Optical symbolic substitution: edge detection using Prewitt, Sobel, and Roberts operators," Appl. Opt. 28, 4644- (1989)

**Nael Hirzallah** Dr. Nael Hirzallah has finished his Masters and PhD Degrees from the University of Ottawa, Canada, in 1993 and 1997, respectively. He then worked in the industry in leading software and semiconductor companies in both Canada and USA. He joined Applied Science University in Jordan in 2002 to work as an Assistant Professor in the faculty of IT. In 2007 he was promoted to Associate. He became its Computer Center Director in 2003. For three years starting 2008, he worked as a department chairman and Acting Dean in Fahad Bin Sultan University in Tabuk, KSA. Since 2011, he returned to Applied Science University to assume the position of the Faculty of Information Technology Dean. His research interests are in eLearning, eCommerce, Multimedia and image processing.

# Rigorous Description Of Design Components Functionality: An Approach Based Contract

Abdelhafid Zitouni<sup>1</sup>

<sup>1</sup> *Laboratory LIRE, Computer Science Department  
Mentouri University of Constantine, Algeria*

## Abstract

Current models for software components have made component-based software engineering practical. However, these models are limited in the sense that their support for the characterization/specification of design components primarily deals with syntactic issues. To avoid mismatch and misuse of components, more comprehensive specification of software components is required,

In this paper, we present a contract-based approach to analyze and model the both aspects (functional and non-functional) properties of design components and their composition in order to detect and correct composition errors. This approach permits to characterize the structural, interface and behavioural aspects of design component.

To enable this we present a pattern contract language that captures the structural and behavioral requirements associated with a range of patterns, as well as the system properties that are guaranteed as a result. In addition, we propose the use of the LOTOS language as an ADL for formalizing these aspects. We illustrate the approach by applying it to a standard design pattern.

**Keywords:** *Architecture Description Language, Design by contract, Design components, Design patterns, LOTOS.*

## 1. Introduction

Component-based approaches have been proposed to create and deploy software systems assembled from components. The use of previously developed components should lead to faster time for complex software applications. Therefore, component-based software development is a promising solution to some of the problems that designers, developers and integrators face when building their systems [6]. Software patterns are a design paradigm used to solve problems that arise when developing software within a particular context. Patterns capture the static and dynamic structure and the collaboration among the different components in a software design. Since a design pattern is a recurring piece of software design, it can be seen as a component, called a design component in [15], and used to reify good design practice from conceptual design building blocks into a composable form. Design components focus on

component-based problem solving instead of component-based implementation.

The benefits of design patterns are that they serve as guidance to the novice designer, and they provide an extended vocabulary for documenting software design. Unfortunately, the descriptive format popularized by these catalogs is inherently imprecise. As a consequence, it is unclear when a pattern has been applied correctly, or what can be concluded about a system implemented using a particular pattern.

In order to address the ambiguity issues associated with design pattern descriptions, we introduce the concept of a design pattern contract as a formalism for precisely specifying design patterns. The responsibility of a pattern contract precisely characterizes the requirements that must be satisfied by the designer when applying a particular pattern.

Formal specification and verification techniques are useful for design analysis in that the sense are more precise, expressive, and unambiguous than the informal ones, such as graphical and textual notations. We argue that in order to achieve effective reuse it is important to specify both functional and architectural properties of a component in terms of formal specifications. Formal specifications are amenable to automation in analyzing component properties and thus facilitate the determination of reuse.

The formal description technique LOTOS (Language of Temporal Ordering Specifications) [4] was originally designed to specify the interactions among communicating processes, thus making it suitable for capturing the architectural (interaction) properties of components.

A contribution of this paper is to provide a rigorous description of component functionality. This description can be achieved by means of contracts [18], using pre- and post-conditions for describing the semantics of component's services. Another contribution of this paper is a proposition of a novel Architecture Description Language (LOTOS-ADL) that has been designed to

address specification of structural and dynamic architectures.

The rest of this paper is organised as follows. Section 2 introduces design patterns, and presents the concepts and notation of the LOTOS and contract. We present a short overview of our approach in section 3, before the main section –sect4- of this paper, we focus on the abstract specification of a component. Section 5 presents the concepts of LOTOS-ADL. Section 6 illustrate a case study and gives an overview of our environment of validation. Section 7 discusses the related work. Finally the last section concludes the paper and gives directions for future work.

## 2. Background

In this section, we introduce some basic concepts and terminology about design patterns, components, LOTOS and design by contract.

### 2.1 Design Patterns

Design patterns are a design paradigm used to solve problems that arise when developing software within a particular context [10]. Patterns capture the static and dynamic structure and collaboration among the components in a software design. To build software systems, a designer needs to solve many problems. Applying known design patterns to address these problems allows the designer to take advantage of expert design experience documented in each pattern. Although design patterns are not formal in nature, design components that have been inspired by design patterns are amenable for formal modeling and analysis. The focus on design components is important because one of the goals of our work is to detect errors as early as possible in the development process by reasoning about the properties at the design level and reducing the cost of finding and correcting these errors in concrete software components.

### 2.2 LOTOS

LOTOS is a formal description technique based on a combination of CCS [19] and CSP [14]. In LOTOS, a system is seen as a process, possibly consisting of several sub-processes. Likewise a sub-process is a process in itself, and a LOTOS specification describes a system via a hierarchy of process definitions. A process is an entity capable of performing internal, unobservable actions, and of interacting with other processes which form its environment. In that sense, LOTOS implements a black

box paradigm used to develop high level, concise and abstract specifications of complex systems. At some abstraction level, it is possible to express the interactions of a process with its environment without having to describe the internal structure (or implementation) of that process. Process definitions are expressed by the specification of behaviour expressions that are constructed by means of a restricted set of powerful operators making it possible to express behaviours as complex as desired. Basic LOTOS is a subset of LOTOS. The processes interact with each other by pure synchronization without exchanging any value. Fig.1 provides an intuitive illustration of the main Basic LOTOS operators.

Operator	Description	Example
[ ]	Either $P1[a,b]$ or $P2[c,d]$ depending on the environment	$P[a,b,c,d]=P1[a,b][ ] P2[c,d]$
[   ]	Parallel composition without synchronization: $P1[a,b]$ is independent from $P2[c,d]$	$P[a,b,c,d]=P1[a,b][   ] P2[c,d]$
[b]	Parallel composition with synchronization on gate b	$P[a,b,c]=P1[a,b][b] P2[b,c]$
	Parallel composition with synchronization on several gates (b,c,d)	$P[a,b,c,d,e]=P1[a,b,c,d][b,c,d] P2[b,c,d,e]$
hide b in [b]	Parallel composition with synchronization on gate b, moreover where gate b is hidden	$P[a,c]-hide\ b\ in\ P1[a,b][b] P2[b,c]$
>>	Sequential composition $P1[a,b]$ is followed, when $P1$ terminated, by $P2[c,d]$	$P[a,b,c,d]=P1[a,b]>> P2[c,d]$
[>	Disrupt: $P1[a,b]$ may be interrupted at any time before its termination by $P2[c,d]$ .	$P[a,b,c,d]=P1[a,b][> P2[c,d]$
:	Process prefixing by action a	a:P
Stop	Process which cannot communicate with any other process	Stop
Exit	Process which can terminate and then transforms itself into stop	Exit

Fig.2. Basic LOTOS operators [3]

### 2.3 Design by contract

Design by contract is a design approach developed by Meyer [18]. It is used here to provide precise specifications for the functionality of components and to enhance their reliability. According to Meyer, a contract is a collection of assertions that describe precisely what each feature of the component does and does not do. The key assertions in the design by contract technique are of three types: invariants, pre-conditions, and post-conditions. An invariant is a constraint attached to type that must be held true for all instances of the type whenever an operation is not being performed on the instance.

Pre-conditions and post-conditions are assertions attached to an operation of a type. A pre-condition expresses requirements that any call of the operation must satisfy if it is to be correct. A post-condition expresses properties that are ensured in return by the execution of the call.

### 3. Overview of the Approach

In [27] we have presented a systematic approach for a software designer to model and analyze component integration during the design phase, the early planning stage of the software lifecycle. This approach includes a process of representing, specifying, instantiating and integrating design components and analyzing their compositions, which are captured as contracts. The process is illustrated in Fig.2.

This approach allows design components to be reused by making the components description available in a component library. With this approach, the designer can not only model the design component precisely, unambiguously and expressively, but also detect the interactions between components and correct design errors before implementation [26]. As shown in figure 2, our approach begins by four steps: (Analysis, selection, abstract specification and the instantiation steps).

In this article we focus on the abstract specification of the component and the ADL for describing architecture of component-based software, which provide explicit support for specifying components. ADLs are important since they can document component-based architecture early, reason about their properties, and automate their analysis and system generation [12].

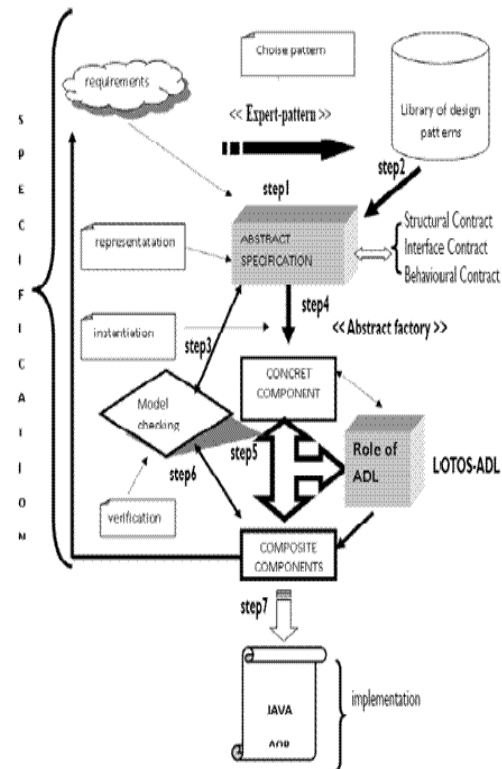


Fig.2. Overview of our approach [27]

### 4. Abstract specification of a component

The abstract specification is inspired from the work of Dong and al. [7]) and contains a formal model of design component, called design component contract. A design component contract includes structural contract (SC), behavioural contract (BC) and interface contract (IC).

The structural properties describe the relations of the constructs of each design component. The behavioural properties are constraints such as event ordering, and action sequence of each design component. The interface contract describes the finite set of input or output ports attached to a design component and the set of messages sent to or received by a component. We define an abstract specification contract (ASC) as:

**ASC ::= <<Component-Name>Where<assertion>and  
 <SC>and<IC>and <BC> End**

### 4.1 A motivating example

To motivate this paper we consider the structure (class and interaction diagram) of the Observer pattern shown in fig. 3 [10]: (The Observer pattern (also called Publisher-Subscriber) regulates how a change in one object can be reflected in an unspecified number of dependant objects).

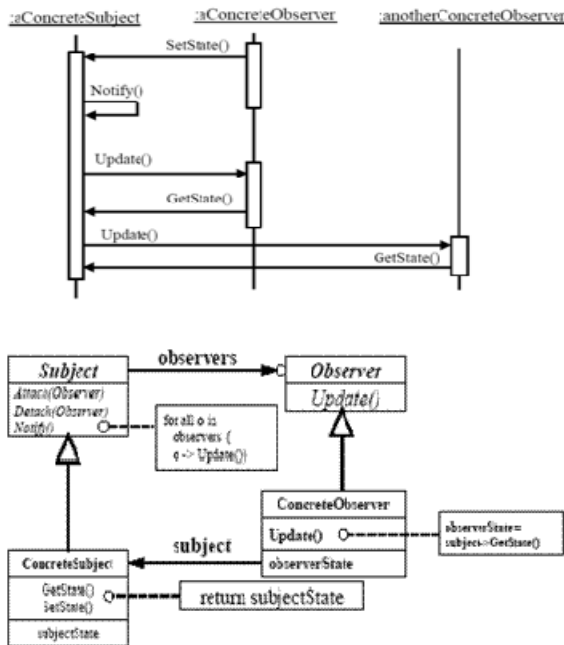


Fig. 3 Observer pattern (class diagram ,interaction diagram)

### 4.2 Structural contracts

In [27] we have formalized the structural aspect of a design component contract by using a subset of First Order Logic (FOL), because the relations between pattern participants can be easily expressed as predicates. The subset of FOL used to describe the structural aspect of a design component comprises variable symbols, connectives ( $\wedge$ ), quantifiers ( $\exists$ ), element ( $\epsilon$ ) and predicate symbols acting upon variable symbols. The variable symbols represent class, objects, while the predicate symbols represent permanent relation [24].

We define two groups of predicates, entities (Table 1) and relationships (Table 2).

- Entity predicates define whether a design component has a specific class (abstract or concrete), what a method (or attribute) is defined in a class....
- Relationship predicates define the relations between classes, attributes, and operations and the actions that a role can perform in a component.

The Abstract specification of a component presented in this paper is an extension of the model existing domain in [27] by introducing the concepts related to the Aspect-Oriented Approach .

For the concepts related to the Aspect-Oriented approach, we will define new predicates. (Table 1) (Table 2)

An Aspect-Class Diagram contains the classes, the aspects and the interfaces, linked with relations, which include associations, generalization, and realization between classifiers and calls between operations.

Table1: Entity predicates

Predicate	Description
Abstract-class ( C )	C plays the role as an abstract-class in the component
Abstract-Aspect(A)	C plays the role as an abstract-aspect in the component
Class ( C )	C plays the role as concrete class in the component
Aspect(A)	A plays the role as concrete aspect in the component
$x \in X$	X is an element of set X

Table2: Relationship predicates

Predicate	Description
Inherit (A,B)	B is a subclass of A
Associate (A,B)	A,B are connected with association relation
Aggregate (A,B)	A contain a reference to B
Invoke (A,m1,B,m2)	A method m1 defined in class A calls a method m2 defined in class B
New(A,m,O)	The method m of class A create a new object of type A
Return (A,m,O)	The method m of class A returns an object O of type A
DeclareParent (A,B,C)	B is a sub class of A. This relation is declared in aspect C
call (A,cp,B)	The Pointcut CP of aspect A designates a set of join points of the class B
advice(A,cp ,action)	Advice codes implement the behavior of an aspect A. Several types of action exist: before, around, after returning,

### 4.3 Interface contracts

We define the interface aspect of a design component contract as follow:

Let a tuple  $IC = (P, IP, OP, IM, OM, IMI)$ , where  $P$  is a finite set of process names,  $IP$  is a finite set of input ports attached to a process,  $OP$  is a finite set of output ports attached to a process,  $IM$  is a finite set of input messages sent to a process and  $OM$  is a finite set of output messages sent from a process,  $IMI$  is the finite set of input messages sent from outside the design component to a process.

The abstract specification of the interface contract of Observer is done by:

- (0) Component-name is Observer where:
- (1)  $\exists (aConcreteSubject, aConcreteObserver, anotherConcreteObserver) \in C$
- (2)  $\wedge \exists (inOS, inSO, self, input) \in IP$
- (3)  $\wedge \exists (outOS, outSO, output) \in OP$
- (4)  $\wedge \exists (attach, detach, getstate, setstate, update, notify, change) \in IM$
- (5)  $\wedge \exists (attach, detach, getstate, setstate, update, notify) \in OM$
- (6)  $\wedge \exists (change) \in IMI$

In order to be able to support dynamic reconfiguration of the service and to provide precise specification about the relationships of operations calls to each other, we include the constraints on component interfaces.

This allows assertions about the gates (set of input or output ports attached to a process) to appear in pre-conditions, and post-conditions.

Let  $IC1 = (IC, Constraint)$  we denote:  
 $p \in IP(p) = \{i \in IP \mid gate\_Ini = p\} \wedge$   
 $p \in OP(p) = \{i \in OP \mid gate\_Outi = p\} \wedge$   
 $m \in IM(p) = \{i \in IP, m \in IM \mid gate\_Ini ?m\} \wedge$   
 $m \in OM(p) = \{i \in OP, m \in OM \mid gate\_Outi !m\} \wedge$   
 $all\_gateIN = \{ \text{all } IP(p) / p \in \text{Component} \} \wedge$   
 $all\_gateout = \{ \text{all } OP(p) / p \in \text{Component} \} \wedge$

**Where** Constraint /\*constraints on gates\*/:

$\forall i, j \in 1..n \rightarrow gate\_Ini \neq gate\_Inj \wedge$   
 $\forall i, j \in 1..n \rightarrow gate\_Outi \neq gate\_Outj \wedge$   
 $\forall i \in 1..n \rightarrow \exists ! i \in 1..n / gate\_Inj ?mi \in gate\_Outi !mi \wedge$   
 $\forall i \in 1..n \rightarrow \exists ! i \in 1..n / gate\_Outj ?mi \in gate\_Ini !mi$

### 4.4 Behavioural contracts

In contrast to the structural aspect of a design component contract, the behavioural contract describes the dynamic information, such as the collaboration among the objects participating in the component and the creation of new objects.

We have chosen a basic LOTOS for defining a formal semantic model of behavioural contracts because it represents a powerful approach for modeling behaviour and concurrency. The choice of LOTOS is motivated by its powerful ability for describing behaviour and the availability of tools enabling formal verification and automatic generation of distributed programs. Our proposal focuses on formally describing architectures encompassing both the structural and behavioural viewpoints. The LOTOS specification of the observer follows:

**Specification** Observer [input,output] : **noexit**:=  
 /\*... Signature.....\*/

**behaviour**  
 $aConcreteSubject [input, output]$   
 $[[input, output]]$   
 $aConcreteObserver [input, output]$   
 $[ ]$   
 $anotherConcreteObserver [input, output]$

**where**  
**Process**  $aConcreteSubject [inCS, outCS] := noexit$   
 $?setstate; !notify; !update; ?getsate;$   
 $aConcreteSubject [inCS, outCS]$

**Endprocess**  
**Process**  $aConcreteObserver [inaCO, outaCO] := noexit$   
 $!; !setstate; ?update; !getstate$   
 $aConcreteObserver [inaCO, outaCO]$

**Endprocess**  
**Process**  $anotherConcreteObserver[inbCO, outbCO] := noexit$   
 $!; !setstate; ?update; !getstate$   
 $anotherConcreteObserver [inbCO, outbCO]$

**Endprocess**

**Endspec**

## 5. Proposal Architecture Description Language

A key aspect of the design of any software system is its architecture. From a runtime perspective, an architecture description should provide a formal specification of the architecture in terms of components and connectors and how they are composed together. Enabling specification



of dynamic architectures is a large challenge for an Architecture Description Language (ADL). This section describes LOTOS-ADL, our proposal ADL that has been designed to address specification of structural and dynamic architectures. While most ADLs focus on defining software architectures from a structural viewpoint, our proposal LOTOS-ADL focuses on formally describing architectures encompassing both the structural and behavioural viewpoints.

From a runtime perspective, two viewpoints are frequently used in software architecture: the structural viewpoint and the behavioural one.

The structural viewpoint may be specified in terms of: components, connectors, and configurations of components and connectors.

**<LOTOS-ADL>:=** < structural viewpoint, behavioural viewpoint>;  
 < structural viewpoint> := <component, connector, configuration>/  
**component** := <cp1, cp2, ....., cpn>  $n \geq 2$  and  
**connector** := <ct1, ct2, ....., ct<sub>m</sub>>  $m \geq 1$   
 with **constraints**:  
 $\forall cp1, cp2 \in \text{component} / \text{name.cp1} \neq \text{name.cp2}$   
 $\forall ct1, ct2 \in \text{connector} / \text{name.ct1} \neq \text{name.ct2}$   
**configuration** := < /\* LOTOS operators construct \*/ >  
 <behavioural viewpoint>:= < LOTOS behavior expression >

Thereby, from a structural viewpoint, an architecture description should provide a formal specification of the architecture in terms of components and connectors and how they are composed together. Further, in the case of a dynamic architecture, it must provide a specification of how its components and connectors can change at runtime. The behavioural viewpoint may be specified in terms of: actions a system executes or participates in, relations among actions to specify behaviours, and behaviours of components and connectors, and how they interact. A LOTOS specification describes a system through a hierarchy of active components, or processes. A process is an entity able to realize non-observable internal actions, and also interact with others processes through externally observable actions.

We model a component as a black-box with a set of input and output gates (or channels), where visible events occur. Instead of describing the static functionalities that a component provides, we specify the set of (dynamic)

behaviors that a component may exhibit in constituting a system. All gates, together with constraints that may be imposed upon the ports, constitute the interface of a component. The interface of a component specifies the constraints on the way the component is to be used. A component may have overall constraints imposed upon the gate. The set of concepts that are manipulated are presented within our ADL meta-model (Fig. 4).

In our meta-model, we are mainly interested in representing static and dynamic behaviour contract using static and dynamic contract. A major benefit of separate static part from the dynamic part is that reasoning independently from any particular situations. The static contract of a component is a part that does not evolve. The evolution of a dynamic contract may have different purposes.

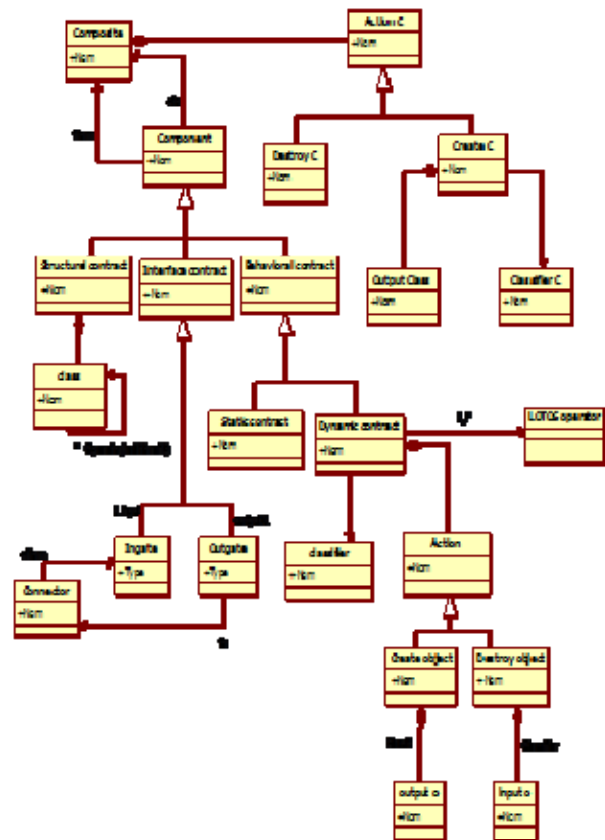


Fig.4. The LOTOS-ADL Meta-model

## 6. Case study: Client/server

Let us, consider the simple client-server system shown in Figure 5. It consists of one client and one server interacting via link connector. Such a system is easy to describe in LOTOS-ADL. A LOTOS-ADL specification describes a system through a hierarchy of components (process). A process is an entity able to realise non-observable actions, and also interact with others process through externally observable actions.

The LOTOS specification at the top-level is a parallel composition of the process Client (component client), the process Server (component server) and the process connector (connector) (Fig.5). In order to specify this system, we adopt the following guiding [22]:

- The basic architecture elements, namely basic components and connectors, are modelled through the basic LOTOS abstraction, namely process.
- Any two LOTOS processes that model components must be in parallel composition with a LOTOS process defined as a connector
- The service specification consists of the temporal ordering of events executed at the service interface.
- We call to invocation (inv) those actions to activate the service and termination (ter) to the action of return a result.

### 6.1. Point to point connector

The LOTOS specification at the top-level is a parallel composition of the process Client (component client), the process Server (component server) and the process connector (connector) (Fig.5).

```

specification Client-Server [invClt,terClt,invSrv,terSrv]
: noexit=
  library RESULT, SERVICES endlib
  behaviour
  Client [invClt, terClt]
    |[invClt, terClt]|
  connector [invClt, terClt, invSrv, terSrv]
    |[invSrv, terSrv]|
  Server [invSrv, terSrv]
  where
  .....
  .....
  Endprocess
    
```

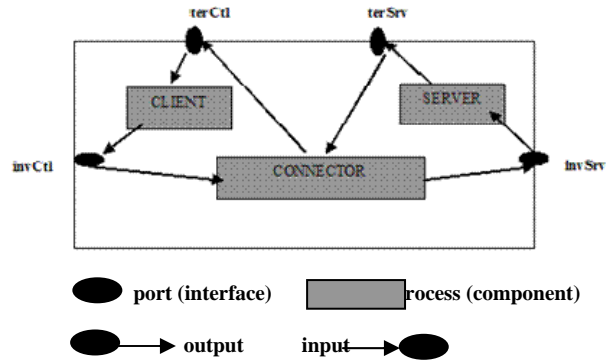


Fig. 5. Illustration of the Client-Server specification

The connector behaviour is defined through the temporal ordering of invocation operations in the connector interface. The connector interface is made up of four ports: invClt to invocations from client, terClt to returns to client, invSrv to invocations from server and terSrv to return to server

```

process Connector[invClt,terClt,invSrv,terSrv] : noexit =
  invClt ? s : SERVICE ? op: OPER /* the client passes the
  request to connector*/
  invSrv ! s ! op; /* the connector passes the request to the
  server*/
  terSrv ! s ? r : RESULT; /*the server passes the reply to the
  connector*/
  terClt ! s ! r; /*the connector passes the reply to the client*/
  Connector [invClt, terClt, invSrv,terSrv]
  Endproc
    
```

In this case, the connector receives an invocation from the server that contains both the name of the requested service and the operation being requested on the server (invClt?s: SERVICE? Op: OPER). The connector passes both of them to the server and waits for the reply. Finally, the connector passes the reply containing the result to the client.

### 6.1. Multicast connector

The connector abstract software architecture is defined as a collection of services. In order to specify the connector abstract software architecture, we assume that is composed by three components (Fig. 6) (service1, service2, service3) and a single connector (communication Service). The LOTOS specification of this software architecture is done by a parallel composition of the set of basic services and the process Communication Service.

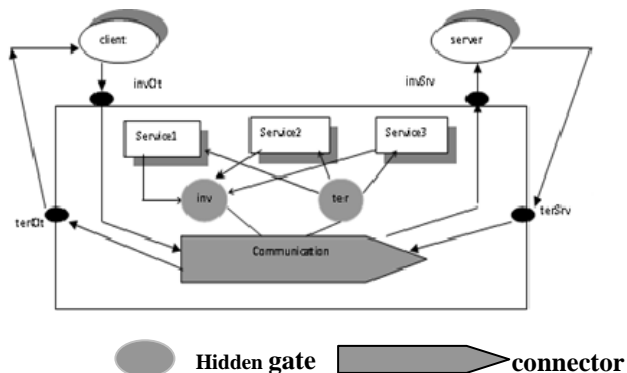


Fig. 6. Illustration of the abstract software architecture

```

ProcessConnector_Abstract[invClT, terClT, invSrv, terSrv] :
noexit :=
  hide inv, ter in
  ((Service1 [inv, ter] |||
   Service2 [inv, ter] ||| Service3 [inv, ter ])
   ||
   ServiceOrdering [inv, ter ])
  |[inv,ter]|
  CommunicationService[inv,ter,invClT,terClT,invSrv,terSrv]
  Where
  .....
  .....
  Endproc
    
```

According to the constraints imposed by ServiceOrdering, after the request gets in the connector, it is passed to Service1 followed by Service2 and Service3. The LOTOS specification of the ServiceOrdering is done by:

```

Process ServiceOrdering [inv,ter] : noexit :=
  inv ! Service1 ? op: OPER
  ter ! Service1 ? r : RESULT
  inv ! Service2 ? op: OPER
  ter ! Service2 ? r : RESULT
  inv ! Service3 ? op: OPER
  ter ! Service3 ? r : RESULT
  ServiceOrdering [invClT, terClT, invSrv,terSrv]
  Endproc
    
```

## 6. Verification

For the verification of our approach, we use our environment of verification, named FOCOVE (Formal Concurrency Verification Environment) [27] (available in [www.focove.new.fr](http://www.focove.new.fr)) (Fig. 7). Focove is an integrated environment designed to edit Basic LOTOS behavior expressions which describe reactive systems and to generate and analyze Maximality based Labelled Transitions Systems structures (MLTS). concerns the state of the art of ADLs.

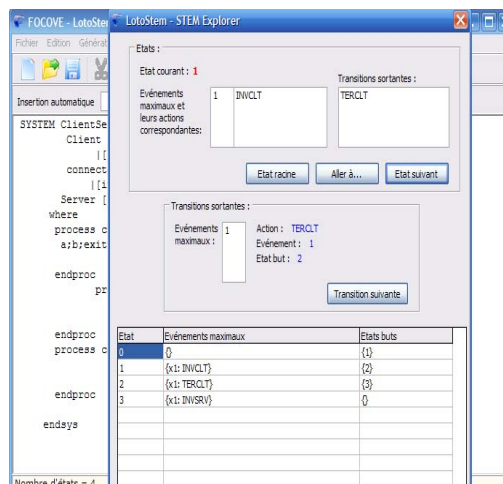


Fig 7. The environment of verification

The FOCOVE environment is dedicated to the design and verification for component based software development. FOCOVE translates a LOTOS program into a Labelled Transition System (LTS for short) describing its exhaustive behaviour. This LTS can be represented either explicitly as a set of states and transitions or implicitly as a library of C functions allowing us to execute the program behaviour in a controlled way. By verification, we mean comparison of a complex system against a set of properties characterizing the intended functioning of the system (for instance, deadlock freedom, mutual exclusion, etc.).

## 7. Related works

We have chosen three dimensions to compare our approach with other existing work. The first dimension concerns the use the design pattern in designing applications. The second dimension concern of application of the contract notion and the approach to software component specification. The third dimension concerns the state of the art of ADLs.

-Several successful experiences have reported on the advantages of using patterns in designing applications [11], [23]. These experiences do not follow a systematic method to develop applications using patterns. Systematic development using patterns utilizes a composition mechanism to glue patterns together at the design level. Generally, we categorize composition mechanisms as behavioural and structural compositions. Behavioural composition approaches are concerned with objects as

elements that play several roles in various patterns. Reenskaug [20] developed the Object Oriented Role Analysis and Software Synthesis method. The method uses a role model that abstracts the traditional object model. Riehle [21] uses role diagrams for pattern composition. The approach by Jan Bosch [4] uses design patterns and frameworks as architectural fragments. Each fragment is composed of roles and components that are merged with other roles to produce application designs.

-The notion of contracts in software development is attributed to Meyer [18]. Another contribution, the object-oriented contracts of Helm et al. [13] focused on specifying the behaviour and interactions between objects in a system. Helm et al. noticed that the behaviour of an object could not be inferred from its interface, leading to design and reuse problems. Contracts formalize the behavioural relationship between objects and define a set of participants and their obligations. In this paper, we defined a formal model of design component based on contracts and a rigorous analysis approach to software design composition.

Keller and Schauer [15] described a methodical approach to design composition which was illustrated as a process within a four-dimensional design space. They characterized a special kind of component, called a design component, and discussed a development process to compose these components at the design level and generate source-code frames or executable code. Although our approach is also in the area of software composition, it focuses on the formal, declarative, and property-based aspects of design composition.

- The majority of the ADLs support only a structural view of the system. Even if offering any techniques for describing behaviour of the system, they only model its possible behaviour and thus can check its consistency only statically (e.g. correctness of proposed configuration, type checking, pre- or post-conditions, protocol). A few of them support dynamic configurations. C2 [25] specifies only pre- and post-conditions, Darwin [17] expresses component semantics in terms of  $\pi$ -calculus. Weaves [12] defines partial ordering of data-flow over input and output objects, but only Rapide [16] and Wright [2] specify dynamic component behaviour. Wright focuses on specifying communication protocols among components and uses a variant of CSP [14] to describe architectural behaviour. It treats both components and connectors as

processes, which synchronise over suitably renamed alphabets. But, it implies a component interface extension in case of permitted reconfiguration and checks only if a connector protocol is deadlock-free as a consistency check [1]. Moreover, none of these ADLs have component have first class in order to cope with description of dynamic and mobile architecture.

#### 4. Conclusions

In this paper, we have introduced a proposition of formal model of design component based on contract and a rigorous analysis approach to software design composition based on automated verification techniques. Our approach allows us to find errors in the design composition early in the development process. This paper has illustrated how to adopt LOTOS as ADL to describe the behaviour of software architecture.

This language is mathematically well-defined and expressive: it allows the description of concurrency, non-determinism, synchronous and asynchronous communications. It supports various levels of abstraction and provides several specification styles. These positive features encouraged us to adopt LOTOS as an ADL for describing both component and connector enables us to check behaviours properties. Finally, LOTOS specifications can also be used to express and verify concurrency models and real-time properties of systems.

The presented LOTOS specifications serve as a basis for very interesting future work. We are currently interested in the refinement of specifications in which the refinement process follows the rules of the software architecture refinement.

Also, we are investigating to proposing a rules-based transformation enabling the mapping from LOTOS specification to JAVA pseudo code.

#### References

- [1] R. Allen, D. Garlan, and R. Douence.: Specifying dynamism in software architectures. In Proceedings of the Workshop on Foundations of Component-Based Software Engineering, Zurich, Switzerland, September 1997.
- [2] R. Allen.: A Formal Approach to Software Architecture. PhD thesis, Carnegie Mellon, School of Computer Science, January 1997.: Issued as CMU Technical Report CMU-CS-97-144.
- [3] Aprille L, Saqui-sannes P, Lohr C.: A new UML profile for reel-time system formal design and validation., in LNCS 2185, 2001

- [4] T.Bolognesi, E.Brinksma. Introduction to the ISO specification language LOTOS. In Van EIJK, pp 23-73, 1989
- [5] J. Bosch.: Specifying Frameworks and Design Patterns as Architecture Fragments. Proceedings of Technology of Object-Oriented Languages and Systems,China, Sept. 22-25 1998.
- [6] Jing Dong.: Design component contracts, Phd thesis. Computer Science department, university of Waterloo, June 2002.
- [7] Dong J, Paulo S C Alencar, Donald D Cowan.: Automating the analyse of design component contracts, In software Practice and Experience, 2005.
- [8] Dong, J., Yang,S., Huynh, D.: Evolving Design Patterns Based on Mode Transformation, *Proceedings of the Ninth IASTED I C S and Applications (SEA)*, pp 344-350, USA 2005.
- [9] Ehrig,H., Mahr,B., Fundamentals of Algebraic specification, volume1, Springer-verlag, Berlin., 1985
- [10] Erich Gamma, Richard Helm, Ralph Johnson, John Vlissides.: Design Patterns, Elements of Reusable Object-Oriented Software, , Addison-Wesley Longman. 1995
- [11] J. Garlow, C. Holmes, T. Mowbary.: Applying Design Patterns in UML. Rose Architect, Vol 1, No. 2, Winter 1999
- [12] M.M. Gorlick, R. Razouk.: Using Weaves for software construction and analysis. In Proceedings of the 13th international conference on Software engineering, pages 23-34. IEEE Computer Society Press, 1991.
- [13] R. Helm, I.M Holland, D. Gangopadhyay.: Contracts: Specifying behavioral compositions in object-oriented systems.Proceedings of the ACM Conference on Object-Oriented Programming Systems, Languages and Applications (OOPSLA),October 1999;
- [14] T. Hoare.: Communicating Sequential Processes. Prentice Hall International, 1985.
- [15] R. K. Keller, R. Schauer. Design Components: Towards Software Composition at the Design Level. *Proceedings of the 20th International Conference on Software Engineering*, pages 302-311, 1998.
- [16] D.C. Luckham, J. Vera.: An event based architecture definition language. IEEE Transactions on Software Engineering, 21(9):717-734, September 1995.
- [17] J. Magee, N. Dulay, S. Eisenbach, and J. Kramer.: Specifying distributed software architectures. In Proc. of 5th European Software Engineering Conference(ESEC'95),,pages 137-153. Springer-Verlag, September 1995.
- [18] Meyer B.: Applying 'design by contract'. IEEE Computer:40-51, October 1992
- [19] R. Milner.: Communication and Concurrency, Prentice Hall, Englewood Cliffs, NJ, 1989.
- [20] T. Reenskaug.: OORASS: Seamless Support for the Creation and Maintenance of Object Oriented Systems. Journal of Object Oriented Programming, 5(6):27-41, October1992.
- [21] D.Riehle.: Composite Design Patterns. Proceedings of Object-Oriented Programming, Systems, Languages and Applications,OOPSLA'97, pp218-228, Atlanta, October 1997.
- [22] N S Rosa, Paulo R Cunha,: A software architecture-based approach for formalising middleware behavior in ENTCS 2004.
- [23] S. Srinivasan , J. Vergo.: Object-Oriented Reuse: Experience in Developing a Framework for Speech Recognition Applications. Proceedings of 20<sup>th</sup> International Conference on Software Engineering,ICSE'98, pp322-330, Kyoto, Japan, April 19-25, 1998.
- [24] Taibi T. and Ngo D.C.L .: Modeling of distributed objects computing design patterns combination. Journal AMCS vol 13 N° 2 pp 239-253, 2004
- [25] R.N. Taylor, N. Medvidovic, K.M. Anderson, E.J. Whitehead Jr., J.E. Robbins, K.A. Nies, P. Oreizy, D.L. Dubrow.: A component and message based architectural style for GUI software. IEEE Transactions on Software Engineering 22(6):390-406, 1996.
- [26] A. Zitouni .: Un framework pour l'utilisation des design patterns par intégration du langage de spécification LOTOS, Congrè International en Informatique Appliquée CIAA05, novembre 2005, BBA, Algérie, ISBN: 9947-0-1042-2
- [27] Zitouni, A., Seinturier, L., Boufaïda.,M., 2008. Contract-based approach to analyze software components, International Conference on *Engineering of Complex Computer Systems* (ICECCS 2008/UML&AADL) workshop, Belfast April, pp 237, 242.
- [28] A. Zitouni, M. Boufaïda, L. Seinturier, "Specifying Components With Compositional Patterns, LOTOS and design by contract". ISCA, 19<sup>th</sup> International Conference on Software Engineering and Data Engineering (SEDE-2010), , San Francisco, California, USA, pp 190-195. June 16 - 18, 2010

**Author** Dr. Abdelhafid Zitouni received his PhD in computer science in 2008 from the University Mentouri of Constantine, Algeria. Currently working as Assistant professor in Mentouri University of Constantine. His research interests include Software Engineering, Software Design, Software Reuse and Design pattern Detection, formal methods in software.

# Method of operational diagnostic state of flow and calculation of calibration Coefficients using artificial neural networks

Dr. Safarini Osama

IT Department  
University of Tabuk,  
Tabuk, KSA

**Abstract**— An important task of operational management in oil and gas production is the control of fluid flow and technological objects of engineering network (EN).

This paper proposed a method for diagnosis of fluid flow measurement and calibration operations. The method is based on the relationship between various parameters of the flow of Engineering Network.

To calculate the actual flow rate on other parameters of the flow, such as pressure, temperature, and the parameters that determine the composition of the liquid (oil), it is proposed to use a neural network.

**Keywords:** *Engineering Network (EN), Neural Network (NN)*

## I. Introduction

An important task of operational management in oil and gas production is the control of fluid flow and technological objects of engineering network (EN). A sign of deviation from the normal operation of EN is the appearance of an imbalance in the nodes of the EN that is mismatch between the amount of expenditure of fluid input and output node.

The causes of the imbalance may be:

- mismatch model EN of the real EN;
- changes in technological regime;
- leakage of fluid;
- uncertainty of the measurement results of liquid flow.

Using the well-known mechanical-mathematical models of processes in hydraulic circuits, and based on these models and methods of analysis of parameters identification of EN oil and gas production difficult because of the complexity of the processes themselves, the incompleteness of the data collected real-time and the large number of unaccounted factors.

In such cases, are becoming increasingly important, methods based on knowledge, which use additional a priori and experimental data on the operation of EN violations in its work. One of the most promising among these methods is neural networks methods [1].

In [1] considered the solution of EN diagnostics with neural networks (NN). In addition to determining the causes of imbalance in our method allows to calculate calibration coefficients measuring tools. The structure of the EN is in the form of the Neural Network, i.e., NN is a model of engineering networks. The weights of the Neural Network are the coefficients that characterize the relationship between the imbalance and the resulting flow measurement (e.g., calibration coefficients of the measurement).

The learning process of the Neural Network with the determination of model parameters on the EN nodes subsequent interval estimation (calibration factors of measurement, the parameters of production lines and inflow / leakage at the nodes) is performed based on the values of a set of vectors imbalances, and the parameters of production lines and inflow / leakage obtained the previous interval estimation. An Education adjustable coefficient of neural network model is repeated until a balance of the corrected measurements of parameters at the nodes of the EN.

From the values of the coefficients obtained by learning, diagnose the state of the equipment, put forward the hypothesis of change in the structure of the site or EN of measuring process parameters, change the initial organizing neural network model, taking into account the indicators of degree of confidence in the models of EN nodes and the results of flux measurements, and repeat the learning process.

Thus, in this method each time the diagnosis is required to conduct its training to determine the calibration coefficients.

## II. Discussion

### Problem statement of operational calibration of the flow rate of EN

To calibrate the instruments necessary to satisfy the following conditions:

- imbalance in the node exceeds the value of EN;
- EN model corresponds to the real EN;
- during the analysis of imbalances no change in the technological regime;

- At the sites corresponding to the edges of the model - the EN is not revealed leakage of the fluid.

These conditions suggest that the cause of the imbalance was the unreliability of the results of measurement of liquid flow. It is therefore necessary to calibrate the flow measuring means of the flow.

To solve this problem, each node IP is considered separately, i.e., without interaction with other nodes. In order to conduct an operational calibration, you must solve the following problems:

1. To determine which edges of the EN unit show unreliable results.
2. Calculate the correction factors to the measured values.

During calibration of the EN is required to find the coefficient  $k$ , which is necessary to multiply the measured flow rate to obtain the real value of liquid flow:

$$Q_r = Q_m k, \quad (1)$$

where  $Q_r$  - the real value of consumption;  $Q_m$  - the measured flow rate.

From equation (1) that the coefficient  $k$  can be calculated from the actual flow rate is:

$$k = \frac{Q_r}{Q_m} \quad (2)$$

Knowing the actual and calculated flow rates you can determine which flows measurements are unreliable. If these two flows close to value, it can be concluded that the results measurements are reliable. In the opposite case it is decided that the results of measurements are unreliable.

Thus, the problem of operational calibration consumption can be reduced to the calculation of the real value of liquid flow.

## Neural network method for calculating correction coefficients

At any time, the flow state is characterized by a number of parameters  $\{P, Q, W, V_{pro}, V_{cr}, C, M, \rho, \eta\}$ :

pressure  $P$ , flow rate  $Q$  (volume or mass), the water content  $W$ , content of dissolved gas  $V_{pro}$ , content of free gas  $V_{cr}$ , content of salts in anhydrous oil  $C$ , content of mechanical impurities  $M$ , density  $\rho$ , viscosity  $\eta$

Between flow and other parameters there is a connection, which is derived from the laws of hydraulics:

$$Q = f(P, W, V_{pro}, V_{cr}, C, M, \rho, \eta). \quad (3)$$

The main idea of solving the problem lies in the fact that the known parameters to calculate the flow rate and compare it with the measured flow rate.

If the calculated and measured costs are equal or close by value, it is decided that the result flow measurement in this thread is valid, i.e., the rate of EN function is correct. If these costs are not equal, then the decision is made that the measurement is invalid.

In this case, knowing the measured and calculated costs we can calculate the correction coefficient of flow measurement this stream. In other words, knowing the other parameters of the flow, we can calculate the flow rate.

One of the most famous formulas that determine the function (3) is the Poiseuille formula [2].

However, its use for the calculation of the other parameters can give a large error, and the calculation is incorrect

There are the following general deficiencies using the analytical formula:

- it focuses on the ideal fluid;
- takes into account not all parameters of each stream;
- contains parameters that cannot be measured for real objects;
- Some of the coefficients in the formula flow cannot be measured for calculations they assumed to be constant (quality pipes, internal friction, etc.).

To solve the problem of calculating the flow rate by indirect methods proposed to use a neural network. In contrast to the formula, NN calculates flow rate, taking into account the characteristics of each particular stream, which is achieved by setting the NN for each thread and its proper training on the actual results of the measurement of various parameters of the flow.

The essence of the Neural Network is as follows:

- for each stream node is assigned its NN;
- NN is trained as an imbalance in the node is less allowable values, i.e., the node is in the normal state;
- NN goes into operation, expects to consumption, when an imbalance in the site than the maximum value and accurately determined that the cause imbalance is the uncertainty of measurement results.

Education of the Neural Network can occur in two ways:

1. If there is a large database of already accumulated information about the values of flow parameters at the same time, whereas the NN can be trained in advance, and as new sets in up to be trained online.

$C, \text{mg}/\text{dm}^3$	$M, \%$	$V_{c2}, \%$	$V_{p20}, \%$	$\rho, \text{MM}^2/\text{s}$	$\eta, \text{kg}/\text{M}^3$	$W, \%$	$D_p$	$Q, \text{M}^3/\text{h}$	$Q_{NN}, \text{M}^3/\text{h}$	$Er, \%$
1000	0,96	0,96	0,94	50	700	0,98	8,2	561,298	561,299	0,00029
1100	0,96	0,96	0,94	51	700	0,98	8,2	550,21	565,18	2,7
1100	0,96	0,96	0,94	51	700	0,98	8,3	556,92	571,59	2,6
1100	0,96	0,96	0,94	51	710	0,98	8,3	549,09	554,71	1,0
1100	0,97	0,96	0,94	51	710	0,98	8,3	554,81	557,47	0,4
1100	0,97	0,97	0,94	51	710	0,98	8,3	560,59	556,72	0,6
1100	0,97	0,97	0,95	51	710	0,98	8,3	566,55	556,09	1,8
1100	0,97	0,97	0,95	52	710	0,98	8,3	555,65	556,32	0,1
1100	0,97	0,97	0,95	52	715	0,98	8,3	551,77	546,73	0,9
1100	0,97	0,97	0,95	53	710	0,98	8,3	545,17	556,62	2,1
1100	0,97	0,97	0,95	53	715	0,98	8,3	541,36	546,89	1,0
1100	0,97	0,97	0,95	54	720	0,98	8,3	527,65	536,76	1,7
1100	0,96	0,96	0,94	52	700	0,98	8,3	546,21	572,91	4,9
1100	0,96	0,96	0,94	53	700	0,98	8,3	535,91	574,27	7,2
1100	0,96	0,96	0,94	54	700	0,98	8,3	525,98	575,68	9,4

Table 1 The results of the neural network

2. In online mode, when the imbalance in the node is less than the permissible value, i.e. it is assumed that all measurements are reliable.

Suppose at a particular time  $T_i$  each flow  $j$  EN has the following set, which characterizes its state:  $\{P_{jib}, Q_{jib}, W_{jib}, V_{projib}, V_{crjib}, C_{jib}, M_{jib}, \rho_{jib}, \eta_{jib}\}$ . This set is said that if the values of the flow parameters are equal  $P_{jib}, W_{jib}, V_{projib}, V_{crjib}, C_{jib}, M_{jib}, \rho_{jib}, \eta_{jib}$  respectively, the flow rate is equal to  $Q_{jib}$ .

If at that moment of imbalance in the node is less than the permissible value, this collection describes the normal state of flow. A lot of these sets describes a set of normal states of the flow.

Suppose at a particular time  $T_i$  node is operating normally, i.e., the imbalance is absent, then supplied to the input of the Neural Network set  $P_{jib}, W_{jib}, V_{projib}, V_{crjib}, C_{jib}, M_{jib}, \rho_{jib}, \eta_{jib}$  and the output  $Q_{jib}$ . Having learned on the set of such collections, the Neural Network is able to recognize the normal state of flow. Further, in case of abnormal operation of the node at a moment of time  $T_k$  when the input, is a set of  $\{P_{kib}, Q_{kib}, W_{kib}, V_{projib}, V_{crkib}, C_{kib}, M_{kib}, \rho_{kib}, \eta_{kib}\}$ . NN finds the most similar state (set parameters) and calculates what should be the rate for given values of the parameters.

Thus, the NN studies situations where the flow is functioning normally, with no deviations. In other words, the NA simulates the flow in the normal mode and in case of abnormal operation (the unreliability of the results) to determine what flow rate corresponds to the current values of other parameters.

To assess the possibility of the NN to solve the problem posed, was conducted a preliminary simulation of the NN.

Simulation of the NN was done in MatLab using the tool nntool.

As a training sample sets were selected  $\{P, W, V_{pro}, V_{cr}, C, M, \rho, \eta\}$ .

As the desired output - the corresponding value  $Q$ . Sets represent different combinations of parameter values. By the formula of Poiseuille [2] for each such set was calculated rate  $Q$ . It is important to note that in this case, this formula is only used for obtaining the training set. The total number of sets was 500.

The results of the efficiency of NN is shown in the table 1.

The first line represents one set of training sample. Other lines - these are examples in which the parameters are changed to a small value compared with the training set, i.e., examples of similar parameter values to the training set.

According to the table the following conclusions can be made:

1. NN gives an error of generalization 1 ... 2% for small deviations of the parameters  $V_{pro}, V_{cr}, C, M, D_p$  and by proportional change in the parameters  $\rho, \eta$
2. Further study of the relationship between density, viscosity and output of the NN.

### III. Conclusions

This paper proposed a method for diagnosis of fluid flow measurement and calibration operations. The method is based on the relationship between various parameters of the flow of EN.

To calculate the actual flow rate on other parameters of the flow, such as pressure, temperature, and the parameters that determine the composition of the liquid (oil), it is proposed to use a neural network

For NN training are encouraged to use the accumulated information about the states of flow, and conduct additional training in on-line mode during normal operation flow.

The method allows to determine which streams the measured flow rate results unreliable, and to calculate correction factors to the EN.

Differences between the proposed method to the case considered in [1] are:

- In this method, NN simulates only one node and not the entire EN;
- Calibration coefficients are calculated based on the output of the Neural Network, which is the calculated value of flow.
- when doing the calculations is not required to regenerate the NN by changing the structure of EN;



- to calculate the calibration factor does not require repeated training of the NN, it is only necessary additional training in the process of operation.

### References

1. Neural network technology in solving problems of analysis and diagnostics utilities / Y.I. Zozulya, D.F. Nazipov, R.R. AKHMETZYANOV, A.A. Geltsov // Automation, telemechanization and communications in the oil industry. - Moscow: JSC "VNIIOENG", 2007. - № 4. - pp. 25-31.
2. Virtual science fund, scientific and technical effects "effective physics." Poiseuille flow. URL: <http://www.effects.ru/science/199/index.htm>.
3. Analysis of the balance in the oil and gas engineering company networks: training materials / M.A. Slepian [and others]. - Ufa: Monograph, 2002. – 120p.
4. Uossermen F. Neuro-Computer Technology: Theory and Practice / Trans. in Russian. language. J.A. Zueva, V.A. Tochenova. - 1992.

### AUTHORS PROFILE



**Dr. Osama Ahmad Salim Safarini** had finished his PhD. from The Russian State University of Oil and Gaz Named after J. M. Gudkin, Moscow, 2000, at a Computerized-Control Systems Department. He obtained his BSC and MSC in Engineering and Computing Science from Odessa Polytechnic National State University in Ukraine 1996. He worked in different universities and countries . His research is concentrated on Automation, Simulation and Control in different branches.

# USABILITY of Collaborative Web Surfing Systems in e-Research

Akhtar Ali Jalbani<sup>1</sup>, Aneela Yasmin<sup>2</sup>, Gordhan Das Menghwar<sup>3</sup> and Mukhtiar Memon<sup>4</sup>

<sup>1,3,4</sup>Information Technology Centre, Sindh Agriculture University  
Tandojam, Pakistan

<sup>2</sup>Department of Biotechnology, Sindh Agriculture University  
Tandojam, Pakistan

## Abstract

Software's developed for the e-Research are generic in nature and are used in diverse application context. The Usability evaluation becomes more severe for that type of software's. In this paper, we focus on e-Research application based software on small scale to evaluate the quality assessment in accordance to the usability of the software. The graphical user interfaces are the main artifacts of the Usability evaluation, we proposed a model transformation method for the generation of executable user interfaces for the usability quality assessment of the software used in Collaborative environment for web surfing.

**Keywords:** *Usability, Collaborative Web Surfing, e-Research, Quality assurance, Model transformation, User Interfaces.*

## 1. Introduction

Now a day's e-Research is getting more attention from the researchers by exploring the new scientific possibilities [1]. Most promising example of e-Research is the usage of text corpora in Humanities. E-Research is gaining more importance in less technical research areas also. Hence, the software development for e-Research is more generic or simple to use. Having generality in the software leads to the flexibility and extensions to the exiting software used for e-research. It is more frequent that the users of the software are less technical and feedback from those users may base only criticism, or we can say that the software does not meet their expectations. Switching from one to another software is possible. User of the software always looks for the software which meets their requirements. Hence, we can say that either the usability or understandability has not been properly studied. The collaborative web surfing is also social network where researchers can surf together. The software's still did not get much attention of the research community because of their bad usability. Hence, the quality of the software decreases, which decrease in the usage of that particular software. In this paper, we focus on the usability of the e-

Research software's used for web surfing in a collaborative environment.

The paper is structured as follows: In section 2, we determine types of software that are used for collaborative e-Research. In section 3 the concept of the usability and its practical approach is discussed. Model based approach for user interfaces generation is discussed in section 4. We conclude with an outlook on future research direction in this area in section 5.

## 2. Collaborative Software's in e-Research

According to Carstensen and Schmidt [2] the concept of collaborative software overlaps with computer supported cooperative work (CSCW) [3] for example groupware [4]. The groupware provides a way that how computer can support collaborative activities and their coordination. Most promising application such as email, calendar, text-chats, wiki and bookmarking belongs to the collaborative way of communication in a groupware. Some other applications are more general for example some social software's such as Twitter [5], Facebook [6] and Friendster [7, 8]. The use of collaborative software in the workspace creates the collaborative environment. A collaborative working environment provides a platform to the individuals and cooperative workers by supporting new class of e-Researchers, who work together beyond the geographical locations. There are three main categories of for working in a collaborative environment.

- **Communication:** It refers to unstructured interchange of information for example phone call.
- **Conferencing:** It refers to the interactive work towards shared information. This collaborative way of sharing goals or brain storming can be used in many applications to achieve common goal.

- **Coordination:** It refers to the complex but based on mutual dependent work to achieve common goal. In this case everyone is doing some thing different to achieve that goal.

The main objective of collaboration software used for e-Research to help and facilitates the e-researchers working together over geographic distances. The tools play important roles in communication, collaboration and process of problem solving in real time shared environment where one or more than one researchers are tried to achieve the common goal.

Recent development shows the application of e-Research Infrastructure is increasing globally distributed through Internet. These types of collaboration are now days built upon grid computing software's, which provide benefits to the researchers in shared environment. The grid computing software's help researches to the usage of advanced ICT tools for data analysis, large scale computing resource and high performance visualizations [9].

### 3. Usability and Its Practical Approach

Usability is used to evaluate the quality of the software that influences the handling of and the user's attitude towards a software product. Usability play important role to various aspects of the software for example decisive role for the selection of process when more than one alternative for same application are available [10].

Usability is based on context or it's a context sensitive. The usability used for one application can not be as good as used for another application [11]. The application context includes: Tasks execution, Environment and User. Usability of the software can be evaluated qualitatively or quantitatively that includes effectiveness, efficiency and error rate. These attributes are indirectly measured based on the observation of the people. Error rate can be calculated by counting number of mistake done by the user, during software usage. These mistakes can be of adding invalid entries, these types of mistakes falls into the qualitative measurement of usability. On other hand, quantitative measurements include satisfaction of the user and attractiveness of the software graphical interfaces. These can be obtained through questionnaire, interviews and surveys. ISO 9241-11 [12] is a standard for usability as shown in Figure 1. In which usability quality measure is divided into two sub tasks, Usability is influenced by application context and Usability is evaluated by measures. Figure 1 presents the application tasks in terms of Tasks, environment and user. Tasks are based on goals and steps taken by the user that particular task. Environment relates to the operating system problems and

user relates to the knowledge, experience and other attributes related to the gender. Measure already discussed relates to the qualitative and quantitative measures.

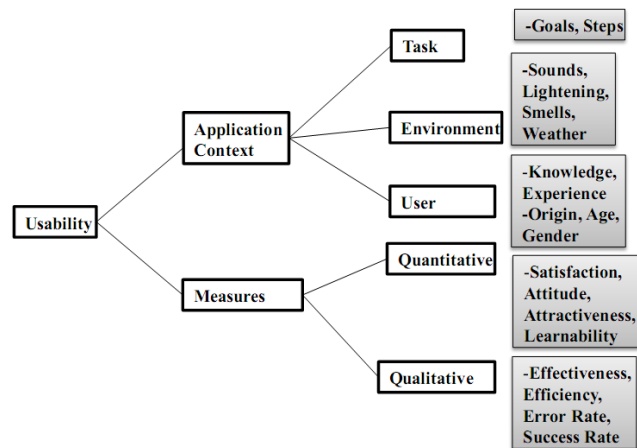


Fig. 1 ISO 9241- 11 Usability measures and its influence.

### 4. Usability for Application Context User Interfaces

User Interfaces and back end are two types of layers used for most of the software's. The user-interfaces can be graphical as well as textual. In this paper only graphical user interfaces are discussed. The graphical user interfaces allow user to work with the software [13]. The back end provides the functionalities implemented for that user interface. The generic or specific software user interfaces can be modeled with three types of the models, these are: 1) User Interface frame work model, which identifies the basic elements of the user interfaces, 2) User Interface model that describes the specific interaction elements and their types and 3) functionality model that describes all functionalities available by specific or generic software.

### 5. Model Example for Collaborative Web Surfing Application

We have developed a collaborative web surfing application called NetSurf [14, 15]. Hence, in this section, we will discuss the practical applicability of usability to its two major component of NetSurf application:

- NetSurf Graphical user Interface (GUI): It provides an integrated environment with web browser to the user to access all functionalities of software.
- Repository at Back-end: in addition to other, it stores and achieves the keyword and their

corresponding URL that enforces the user management including access rights.

On back-end user is unaware of the functionality of the NetSurf in which how data is stored in the repository. The user can access to the GUI buttons, text and video messaging and file transfer functionalities. Each of the user access to the functionalities that are logically belongs together.

As a refinement of the user interface framework model, the user interface for collaborative NetSurf software is application specific software that defines the concrete interfaces of the elements in that context for example data transfer, which defines specific representation of the contents in the NetSurf. It also defines the specific buttons, labels, elements, chat and data transfer buttons and context menus. Furthermore the user interface model links the defined elements of the NetSurf to the appropriate function, in which data is added to the repository of the software. The functionality model can also define the combinations of the functionalities of the functional model to appear as one the functionality.

To assess quality of the software, rules and guidelines play important role. Guidelines are the best practices, which are gathered by the experts with their experience and practical knowledge. The guidelines may also vary from domain to domain. In quality engineering assessment based on usability is divided into three main categories [16].

1. The human behavior and psychology for the application used in the particular domain
2. Guidelines for the application used in particular domain
3. Usability evaluation based iterative user interface

From above quality assessment approach, the most effective assessment for the usability is conducted through iterative evaluation of user interface application. The advantage of this method is that it can be applied at any stage of the software varying from paper-based interface drafts up to final versions of the user interfaces.

However, quality engineering methods applied for the user interfaces or any other applications have two major methods for good usability products. These are expert oriented and user oriented methods. The experts methods are applied early stages of design and user oriented methods are applied in later stages. The disadvantage of expert oriented method is that it does not involve the end user only expertise of the developer is utilized. Hence, one can compare both types of the methods for quality engineering easily. User oriented used in later stages focusing on the functionality and quality of the product has some advantages over the expert method.

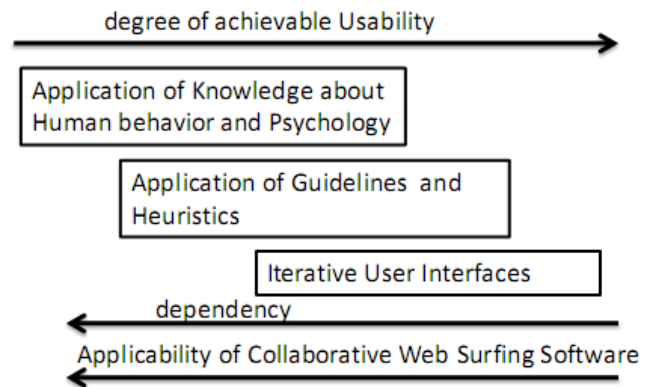


Fig. 2 Application of usability in a collaborative web surfing.

Following relationship is illustrated in the Figure 2, which defines the applicability of the usability for the generic software's. The boxes show the different categories of the usability engineering and arrows indicate the degree to be achieved by usability, dependency and applicability for the software application.

## 6. Model Transformation approach for User Interface Generations

Model based software development is an emerging topic these days. Model based approach can also be used to generate executable user interface model for any application [17, 18]. To use model transformation approach, we used Unified modeling language for the designing of user interfaces, with class diagrams.

Model transformation approach works in the following way [19]. The three types of the user interface model can be served as input models. To get executable user interface model, we need to apply transformation rules on the input models. These transformation rules are applied into XPand model to text transformation language [20]. The XPand language is based on the EMF models; hence the input UML model for user interface is EMF model [21]. The model transformation language is a part of the eclipse modeling project. The output executable model is basically source code for user interface that can be the part of the software. The source code is an executable and can be generated the Graphical user interface for the software. Figure 3 shows the model transformation approach in a generic way. In which Input model are user interface frame work model, user interface model and the functionality model to produce executable model based on the transformation routines.

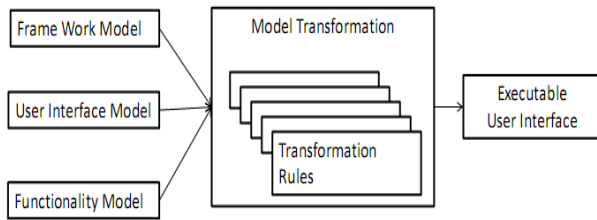


Fig. 3 Model Transformation approach for Executable user interfaces.

## 7. Conclusions

The same usability can not be achieved with the generic software because the usability is based on the application context. In this paper, we provide how usability can be applied on the application based software in the domain of web surfing. Different types of quality engineering approaches have been discussed based on the application context for the user interfaces which has direct impact on the Usability quality attribute and we have showed how model transformation approaches can be applied for executable user interface generation.

Furthermore, our application based software is very small for usability quality assessment. The approach can be applied to the large scale software application based on grid computing for example TextGrid [22]. We can compare their possible results for both applications to see how it is useful to create good quality user interface to improve the usability of the software.

## References

- [1] European Commission, "Community Research and Development Information Service-e-Infrastructure", Accessed on Dec 26<sup>th</sup>, 2011 <http://cordis.europa.eu/fp7/ict/e-infrastructure/>
- [2] P.H Carstensen and K. Shimdt. "Computer Supported Cooperative Work: New Challenges to Systems Design". Handbook of Human Factors, 1999.
- [3] J. Grudin, "Computer-Supported Cooperative Work: History and Focus". Computer 27 (5):1999, 19–26
- [5] Twitter, Accessed on Dec 26<sup>th</sup>, 2011 <http://www.twitter.com>
- [6] Facebook, Accessed on Dec 26<sup>th</sup>, 2011 <http://facebook.com>
- [7] Friendster, Accessed on Dec 26<sup>th</sup>, 2011 <http://friendster.com/>
- [8] D.M Boyd and N.B Ellison "Social Network sites: Definition, History and scholarship" Journal of Computer-Mediated Communication, 13(1): 2007
- [9] J. Goecks, and E.D Mynatt. "Leveraging Social Networks for Information Sharing", Proceedings of the ACM Conference on Computer Supported Cooperative Work, USA, 2004.
- [10] C. Gutwin and S. Greenberg. "The Mechanics of Collaboration: Developing Low Cost Usability Evaluation Methods for Shared Workspaces", Proceedings of the 9th

IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises, Washington, 2000.

- [11] J. Bosch and J. Natalia. "Designing Software Architectures for Usability". Proceedings of the 25<sup>th</sup> International Conference on Software Engineering, 2003.
- [12] ISO. "ISO 9241-11: Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs) -- Part 11: Guidance on usability", 1998. Accessed on Dec 26<sup>th</sup>, 2011 [http://www.iso.org/iso/catalogue\\_detail.htm?csnumber=16883](http://www.iso.org/iso/catalogue_detail.htm?csnumber=16883)
- [13] H. Halpin, V. Robu and H. Shepherd. "The Complex Dynamics of Collaborative Tagging", Proceedings of the 16<sup>th</sup> International Conference on World Wide Web, 2007
- [14] A. A Jalbani, S. Abbasi, G. D Menghwar and A. Yasmin., "Towards an Approach for Web Surfing in Unison". In Proceedings of 4th International Conference on Developments in eSystems Engineering Dubai, UAE 6-8 Dec 2011.
- [15] A. A.A Jalbani, S. Abbasi, G.D Menghwar and A. Yasmin. "Using Collaborative Environment for Web Surfing." Pak. J.Agril., Agril. Engg., Vet. Sci., 2011, 27 (1): 94-99.
- [16] A. Holzinger. "Usability Engineering Methods for Software Developers". Communication of the ACM 48(1), 2005.
- [17] S. Abrahão, E. Iborra, and J. Vanderdonck. "Usability Evaluation of User Interfaces Generated with a Model-Driven Architecture Tool". Information Systems Journal, 2008, pp 3-32.
- [18] H. Traetteberg. "Model-Based User Interface Design". PhD. Thesis, Department of computer and Information science, Norwegian University of science and Technology, Norway, 2002.
- [19] K. Czarnecki and H. Helsen. "Classification of Model Transformation Approaches", proceedings of second OOPSLA'03 Workshop on Generative Techniques in the Context of MDA, USA 2003.
- [20] Iteims. "Xpand Model to Text Transformation Language" Access on Dec 26, 2011. <http://eclipse.org/modeling/m2t/?project=xpand>.
- [21] Eclipse.Org. "EMF- Eclipse Modeling Framework", Accessed on 26 Dec, 2011 <http://eclipse.org/modeling/emf/>.
- [22] TextGrid, "TextGrid".Georg-August-Universitat Goettingen. Accessed on Dec 26<sup>th</sup>, 2011. <http://www.textgrid.de>

**Akhtar Ali Jalbani** is Assistant Professor, Information Technology Centre, Sindh Agriculture University Tandojam Pakistan. He has obtained PhD Computer Science in 2011 from Institute of Computer science, University of Goettingen, Germany. He is a member of System Design Language Forum Society (SDLForum). His research interest includes quality of UML models, Model Transformations and Data Mining.

**Aneela Yasmin** is Assistant Professor, Department of Biotechnology, Sindh Agriculture University Tandojam. She has obtained PhD (Dr. rer. hort. / Molecular Biology) in 2010 from the Institute of genetics, University of Hannover, Germany. Her research interest includes investigating functionality of resistance genes through different techniques of forward/ reverse genetics, computational biology and Data Mining.

**Gordhan Das Menghwar** is Assistant Professor, Information Technology Centre, Sindh Agriculture University Tandojam

Pakistan. He has obtained PhD in Wireless Communications in 2010 from Vienna University of Technology, Vienna Austria. His research interests include cooperative communications, space time codes, network coding and information theory.

**Mukhtiar Memon** is Assistant Professor, Information technology Centre, Sindh Agriculture University Tandojam Pakistan. He has obtained PhD in Software Engineering in 2011 from University of Innsbruck, Austria. He has special interests in modeling security aspects of service-oriented systems. On the technology-side he excels in UML, Model Transformation, WS-Security and Enterprise Integration.

# Search engine optimization with Google

Vinit Kumar Gunjan<sup>1</sup>, Pooja<sup>2</sup>, Monika Kumari<sup>3</sup>, Dr Amit Kumar<sup>4</sup>, Dr (col.) Allam appa rao<sup>5</sup>

<sup>1,2,3</sup>Department of computer science and engineering  
School of engineering & technology, Sharda University  
Greater Noida, GautamBuddhaNagar (U.P)-201306,India

<sup>4</sup>BioAxis DNA Research Centre  
L.B Nagar, Hyderabad, India

<sup>5</sup>JNTU, Kakinada, India

## Abstract

Search engine optimization is a strategical technique to take a web document in top search results of a search engine. Online presence of an organisation is not only an easy way to reach among the target users but it may be profitable too if optimization is done keeping in view of the target users as of the reason that most of the time users search out with the keywords of their use (Say; PhD in web technology) rather than searching the organisation name, and if the page link comes in the top positions then the page may be profitable. This work describes the tweaks of taking the page on top position in Google by increasing the Page rank which may result in the improved visibility and profitable deal for an organisation. Google is most user friendly search engine proved for the Indian users which give user oriented results .In addition, most of other search engines use Google search patterns so we have concentrated on it. So, if a page is optimised in Google it is optimised for most of the search engines.

**Keywords:** Search engine optimisation, SEO, Google optimisation, On page optimisation, Off page optimisation, Image optimisation, URL structure optimisation.

## 1. Introduction

Users use search engines for most of their queries but they only prefer the results available on first page and 2-3% of users go on further pages (except Researchers), Now imagine if the page of an organisation is on 2-3<sup>rd</sup> or 4<sup>th</sup> page then the business which can be generated from that page has a very less change to return and user will prefer the page coming on the 1<sup>st</sup> page. Trillions of web pages are indexed per day in a search engine.

There are millions of search per day .Most of the visitor's visit the website by hitting the links available in search engines and believe that companies found on the top results are the best brand in their product service and category. These clues make it very

clear that if an organisation wants to go on top in their sales then they should concentrate in getting their page widely available in the search engines. For example, if someone wants to use cab services and unknown to the place where he is now, normally if he/she is a techie search of for cab services with the name of city and hit the top 10 links and use their services. There are so many business of online booking system of tickets are growing these days and getting a very good response in very short span of time; in this particular case its very necessary to be on top results of a search engine so that the customers can easily be fetched out.

## 2. Description

### 2.1 Search engine optimization

It is the way of increasing the visibility of a page by natural means i.e., unpaid search results. In this process the website undergoes redevelopment to make our keywords effectively communicate with major search engines. This work is done by SEO (Search engine optimizers), They may target image search, academic search, local search, video search. Optimising a page involves editing contents & HTML codes in order to increase its relevance to specific keywords and proper indexing in search engines .The contents and codings are edited keeping in view of the indexing pattern of the search engines which are done by a crawler named Googlebot in Google. It is the most powerful way to reach to reach the customer as we meet them when they are in need. Most of the users find the target websites during their search.



Figure.1 Search engine optimisation

## 2.2 Page Rank

It is an algorithm used by Google which assigns numerical weight to the URL of web documents to measure its relevance. The numerical weight that it assigns to any given element  $E$  is referred to as the PageRank of  $E$  and denoted by  $PR(E)$  [1]. Stanford University is the birthplace of PageRank when Larry Page (hence the name *Page-Rank*) and Sergey Brin were involved in research of a new kind of search engine. The idea of Sergey Brin was that information on the web could be ordered in a hierarchy by "link popularity": a page is ranked higher as there are more links to it. In 1998, the first paper describing the PageRank and initial prototype was published after which Page and Brin founded Google Inc., the company which is behind the Google search engine. It shows the popularity or a particular link or a website. The page with higher rank gives more optimised results.

$$PR(A) = (1-d) + d (PR(T1)/C(T1) + \dots + PR(Tn)/C(Tn)) \quad [2]$$

The name "PageRank" is a trademark of Google, and the process has been patented (U.S. Patent 6,285,999). The said patent is of Stanford University to which Google has exclusive license rights. The university received 1.8 million shares of Google in exchange for use of the patent; the shares were sold in 2005 for \$336 million [3].

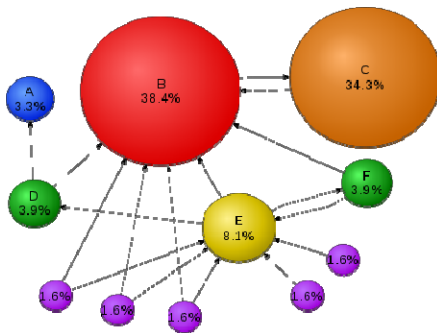


Figure 2: Page ranking in Google

## 2.3 Crawlers & Database

It is a computer programme which browses the World Wide Web in a methodical, automated manner or in a orderly fashion. It normally visits the URL'S of our website [4].

Google	Googlebot
MSN	MSNbot
Yahoo	Yahoo Slurp

Table1: Search engines and their crawlers

## 2.4 Onpage optimisation

It is the first step which every webmaster should concentrate, this deals with the changes we do in our page in order to improve visibility and rank. On Page Optimisation is optimising your website in a way that it can rank better in search engines and improve visitor satisfaction. This optimisation technique depends on nature and business of our website. It is advisable to update the contents of our website and optimise the content each time as these factors are directly related to the content and structure of the website. Modifying Title, Body text, Hyperlinks, URL, Quality and easy to understand contents, increasing the frequency of keyword, robots.txt, sitemaps, Image optimization etc which requires extensive research with the competitor webpages. If proper Onpage optimisation is done, results in drastic increase in the rank and readability of the website[5].

## 2.5 Off page optimisation

This is the work which is done apart from the website to improve the visibility & ranking of a page. Off page search engine optimization is supposedly the complement of On Page Optimization It mainly concentrates in creating backlinks & social media marketing. It is very novel practice to have links from a webpage which has good rank and visibility. It is the best technique to go ahead of the competitors if the webmaster team is equipped with quality of web researchers. In brief it consists of various link building methods like Blog posting, Social networking, Press release, Video submission, link exchange, Article submission etc [6].





Figure 3: Figure showing off page optimisation [7]

## 2.6 Search engine Anatomy

There are four parts in a search engine is observed when a query is done, we may call them as the part of search results. The engine also indicates how many results it has fetched and in how much duration.

### 2.6.1 Non sponsored listing

These are the results from the listings that are done by Google crawlers according to their ranking algorithm. For these listing we need not to pay anything to Google, The pages are ranked according their quality.

### 2.6.2 Sponsored Listing

These results are also shown in the result page at the top most and the right corner, for these listing we need to pay to Google.

### 2.6.3 Search box

This portion is used by the user for his query; it may be from his country or from World Wide Web.

### 2.6.4 Google instant

As we start typing out our query in Google, it starts displaying our result analysing each word. This feature depends on the speed of connection, many a times it doesn't work on slow connection.

## 3. Methodology

### 3.1 On Page Optimisation

#### 3.1.1 Title optimisation

It is a piece of HTML keyword which describes what is website all about to the search engine and users, it is the most important part of a website which is used by search engines to find the relevance of a website. This is the structure how it looks link in coding part. It is advised to use the most relevant keywords in the title tag which describes the website [8].

```
<head>  
<title>SEO India - search engine optimization India, seo  
services, seo company India, affordable seo India  
Chandigarh </title> <head>
```

Following screenshot shows how search engine gives it relevance when fetching out user query.



Figure 4: Figure showing title optimisation

#### 3.1.2 Body text

Contents are the success key for ranking in search engines, so it's important to concentrate on the contents of the website which help contents to be considered by the search engine crawlers at the time of assigning the rankings. Following tweaks have been implemented on our project: -

- 1) Use of heading tags.
- 2) Word frequency: -On an average we had provided 500 to maximum of 800 of words on each page.
- 3) Keyword density: -Frequency of keyword to be optimised was kept 3%-5% on the pages with 500-700 words & 8%-10% on the pages with 700+ keywords.
- 4) Relevant keywords: -Most important keywords of the users query were used carefully specially on the top of page. In general the keywords appearing on the top of a page or top area are most prominent for indexing by the crawler.

### 3.1.3 Hyperlinks

A hyperlink is a navigation element or reference of a document in the other part of the same document, or a specified section of another document, that automatically brings the referred information to the user when the navigation element is selected by the user. The search engines basically predict that if we are linking something from our page is closely related to our page; In brief it makes the contents user-friendly if seen from the search engine point of view. Snapshot of a hyperlink is shown below.

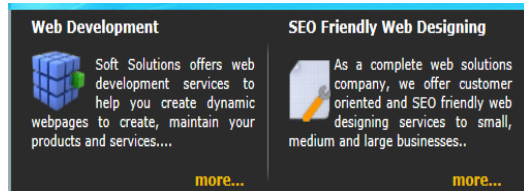


Figure 5: Figure showing hyperlinks

All the words more... are hyperlinks to get detailed information about the respective section.

### 3.1.4 URL

We should improve the structure of url's by using words as simple to understand URL'S will convey content information easily. If our URL'S contain relevant keywords, it provides users and search engine with more information about the page than an ID or oddly named parameter would as the URL to a document is displayed in the search results after the title.

### 3.1.5 Quality and easy to understand contents

Creating and using useful contents increase the influence of a page more than all the tweaks. This tweak is very important in the sense that if a user likes the content then he/she shares it happily via blog, email, forums or other means. We should think from user point of view whether what he searches out to find his contents in a search engine, in addition to it we should create a new and fresh contents ,useful service that no other site offers. Content should be written in a manner that user enjoys the content and it is easy to follow and it should be created in view of users not search engines.

### 3.1.6 Meta Tags optimisation

Meta tags are very useful in providing the search engine about the proper information of a website. Below a complete metatag used by us is shown. Out of all meta tags description tag is most important as it is a part of search results and if the optimisation keywords are provided here properly results with very nice result.

```
<meta name="description" content="BDRC is most trusted Life Science Organisation reco  
operate from Hyderabad Secunderabad Lucknow Bhubaneshwar">  
  
<meta name="keywords" content="Biotechnology Training,Bioinformatics Training ,DNA Te  
<meta name="google-site-verification" content="UUwivXTo0yQeaxXoGemqsiSszumm3Jd653nHd6l  
<meta name="language" content="ENGLISH">  
<meta name="product_brand_name" content="BioAxis DNA Research Centre">  
<meta name="product_family" content="Internet">  
<meta name="region" content="GLOBAL">  
<meta name="distribution" content="Global">  
<meta name="revisit-after" content="15 days">  
<meta name="abstract" content="www.dnares.in">  
<meta name="copyright" content="BioAxis DNA Research Centre,2007-2010">  
<meta name="author" content="BioAxis DNA Research Centre">  
<meta name="robots" content="All">  
<meta name="rating" content="General">
```

Fig 8: Figure showing Meta tag optimisation

### 3.1.7 Newsletters

Many a times when a user visits a website and wants to be updated with updated of the company to which the website belong ,In this case newsletter are the best options ;In this the users provides his/her email-id over there & if there any update comes over the page it is sent automatically sent to the users inbox.

### 3.1.8 robots.txt

This file is used on the files of our website whom we want should not be accessed by the crawler; it is kept in the root directory of the website. If we have some subdomain of our website and want its access to be limited by the web crawler then by creating a robots.txt file for this we may prevent its access to the crawler.

### 3.1.9 Sitemaps

This is a simple page in our website containing the listing of the pages on our site, which displays the structure of our website in a hierarchical way. We should always make two sitemaps, one for users and other for search engines and make the sites easier to navigate. Sitemaps designed for visitors help visitors if they have problems finding the pages

on a site & the sitemaps designed for search engines makes it easier for search engines to discover the pages of a site.

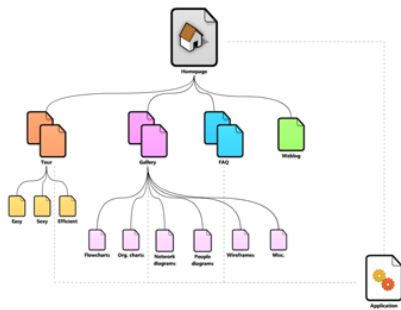


Figure 9: General purpose sitemap

Site map		
<b>Top:</b>	<b>Card category:</b>	<b>Special features:</b>
<ul style="list-style-type: none"> <li>News</li> <li>About this site</li> <li>Privacy policy</li> </ul>	<ul style="list-style-type: none"> <li>By team</li> <li>By players</li> <li>By year</li> <li>By price</li> </ul>	<ul style="list-style-type: none"> <li>Card exchange</li> <li>Bargain pack</li> <li>Holiday gifts</li> </ul>

```

<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  <url>
    <loc>http://www.brandonsbaseballcards.com/</loc>
    <changefreq>daily</changefreq>
    <priority>0.8</priority>
  </url>
  <url>
    <loc>http://www.brandonsbaseballcards.com/news/</loc>
  </url>
  <url>
    <loc>http://www.brandonsbaseballcards.com/news/2008/</loc>
  </url>
  <url>
    <loc>http://www.brandonsbaseballcards.com/news/2009/</loc>
  </url>
  <url>
    <loc>http://www.brandonsbaseballcards.com/news/2010/</loc>
  </url>
</urlset>
    
```

Figure 10: Examples of an HTML site map & XML sitemap. An HTML site map helps users easily find content which they are looking for, and an XML site map helps search engines in finding pages on our site.

### 3.1.10 Hidden target keywords

In this technique we wrote the keywords on the pages by using hyperlinks with other pages full of keywords. Which were being optimised? Keywords which matching colours of the page so that it seems to be invisible to the users were used, These keywords remained from users but it worked a lot as it was accessed by the crawlers.

### 3.1.11 Image optimization

In this technique when optimising a page with targeted keywords we should name the image with the targeted keyword name, it has also a good impact in optimisation. Image search optimization techniques can be viewed as a subset of search engine optimization techniques that focuses on gaining high ranks on image search engine results.

## 4. Off Page optimisation

### 4.1 Backlinks generation

In this process we tend to increase the link of our website on other websites. According to search engines if a page has more and more number of backlinks means there is something relevant in a page because of which other pages are providing links to it, Backlinks are the best way to increase the rank of a page/website and the best way to increase backlinks is link exchange or submitting the URL while doing online advertisements[9].

### 4.2 Blog posting

It is always a good practice to be on a blog like Blogspot,wordpress,thoughts ,linkorbits etc.According to a survey it has been found that companies having blogs are 55% more visitors,97% more inbound links,434% more indexed pages[10].

### 4.3 Social networking

It is the latest technique to have better brand visibility. It is a process sharing information on sites that facilitates content sharing, data exchange, adding unique content etc.Different social media tools includes blogs, podcasts and community based web portals such as Facebook, MySpace, LinkedIn, Twitter, Digg, Reddit etc.These social media tools come with different features like text, images, audio & video sharing among users contents.Following are the benefits of social networking sites: -

- 1) It generates free website traffic
- 2) It boosts up the brand visibility
- 3) It generates inbound links

### 4.4 Press release

Optimizing a press release provides some additional lift to a web site when that press release is distributed and syndicated through other relevant industry or news sites. We shouldn't miss the opportunity to generate valuable backlinks back to our site, driving up our rank and increasing the authority of our site with search engines[11]

### 4.5 Video submission

Videos can be used in several ways to enhance search engine optimisation only the thing is that the videos are

relevant, informative and full of informations. Step by step videos which concentrate on the procedures are best considered. Some of the tricks for video optimisation is given below: -

- 1) Give the video a good title that uses a related key phrase relevant to your product, service or brand.
- 2) Use Video as a pathway to content on your site. Upload videos to YouTube and provide links back to your site.
- 3) Optimize the video for important key phrases using Tags with these terms including even the name of your video.
- 4) Use classic content on the page around your videos with can be indexed by the search engines.
- 5) Keep the videos preferably under 5 minutes but shorter is even better.
- 6) Use a video sitemap with the keywords in the anchor text links so that users and the search engine can find it.
- 7) Tag the videos with key phrases that are relevant to your content.
- 8) Make sure about logo in the video as it will generate brand awareness with your viewers.
- 9) Use the 'Embed Option' when uploading the videos as it allows other users to post the video on their sites/blogs.
- 10) Use descriptive Meta data with relevant keywords and include a keyword rich description of the video.
- 11) Let users rate the video as those with higher ratings tend to be bookmarked and also sent to friends more
- 12) Submit the video using RSS

There are several ways you can now use your videos to enhance your SEO. You need to make sure videos are relevant and informative, providing useful information. Videos that show step by step procedures are excellent as are videos that provide an opinion about a specific topic

#### 4.6 Article submission

Article submission has same impact as of Blog promotion and press release submission.

#### 4.7 Reputation management



Fig 11: Reputation management in search engine optimisation

Search engine reputation management helps to move out of the first result pages those negative posts. It can help to bring back the good name, it helps in keeping business reputation preserved and protected. Each and every corner of the website is monitors and effective measures to protect a good reputation are taken. In brief this service takes effective measures to protect a good reputation and prevents other to damage the reputation.

### 5. Tools Used[12]

#### 5.1 Google webmaster tool

Google Webmaster Tools is a no-charge web service by Google for webmasters. It allows webmasters to check indexing status and optimize visibility of their websites. It has tools that let the webmasters:

- 1) Submit and check a sitemap
- 2) Check and set the crawl rate, and view statistics about how Googlebot accesses a particular site
- 3) Generate and check a robots.txt file. It also helps to discover pages that are blocked in robots.txt by chance.
- 4) List internal and external pages that link to the site
- 5) See what keyword searches on Google led to the site being listed in the SERPs, and the click through rates of such listings
- 6) View statistics about how Google indexes the site, and if it found any errors while doing it
- 7) Set a preferred domain (e.g. prefer example.com over www .example.com or vice versa), which determines how the site URL is displayed in SERPs.

#### 5.2 Meta Tag analyser tool

Following are the uses of Meta Tag analyser tool

- 1) See how search engine robots analyze your or your competitors web site
- 2) Receive tips on how to improve your Meta Tags
- 3) Check the keywords used on the page and find the keyword density

- 4) Check web server operating system where site is hosted
- 5) Check website load time
- 6) Check website file size
- 7) Check URLs and links found on the page

### 5.3 Link popularity check tool

Popularity of a website is checked using this tool. This tool shows how many other sites are linking to the site. Most search engines use this data to calculate how popular your website is. The more links to our site, the better the search engine rankings will be. We can even provide some competitors' URLs to compare our site to theirs.

### 5.4 Sitemap submission tool

This tool is helpful in submitting sitemap to various search engines.

### 5.5 Keyword suggestion tool

This tool suggests keywords related to our keyword which can be used while doing optimisation.

### 5.6 Keyword Traffic estimator

This tool shows us approximately how many daily searches our keywords would get. This tool is used to research the best keywords for our website.

### 5.7 SEO dictionary

It is the list of SEO related keywords with their definitions.

### 5.8 Page rank checker

This tool is used to check the rank of the page so that further actions can be taken to improve the rank of a page.

### 5.9 Page snooper

This tool is used to see the source code of any online site to see the exact structure of the website.

### 5.10 Broken link checker

This tool checks the outgoing links on the page to see if they are broken.

### 5.11 Link counter

This tool counts the number of outgoing links or URL's on a given page and display results. This tool could be useful for link exchange purposes, as we should not trade links with pages with too many outgoing links. It is recommended not to trade links with pages that have over 50 links.

### 5.12 Reciprocal link counter

This tool checks if any given list of sites are linking to your website. It is a great tool to keep track of the reciprocal links to make sure your partner has not removed the link, without visiting their page. You can put up to 100 URL's of sites that you would like to check.

## 6. Conclusions

This paper proposed the novel methods or search engine optimisation for driving more and more users to a website. We used the methods continuously for 24 weeks and discovered more and more users accessing our project website. The rank of the website was raised from 1 to 4; In addition to it sub links were assigned by Google which is assigned to a website which has more number of users according to Google. As a future work we would to develop tools which can add a site to a search engine whenever user wants and can remove the sites which are not good for Mankind. Our project can be visited on <http://www.dnares.in>.

### 5.1 SEO impact Percentage

1	Title	90%
2	Backlinks	75%
3	Domain and file names	75%
4	Description Tags	66%
5	Image optimisation	65%

Table2: Impact of various factors according to SEO point of view.



Figure 12: Figure showing SEO impact percentage

### 5.1.1 Factors of Seo ranking[13]

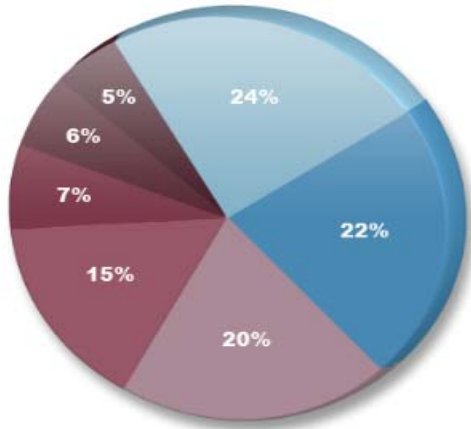


Figure 13: Factors effecting SEO ranking

1. 24% Trust/Authority of the Host Domain
2. 22% Link Popularity of the Specific Page
3. 20% Anchor Text of External Links to the Page
4. 15% On-Page Keyword Usage
5. 7% Visitor/Traffic & Click-Through Data
6. 6% Social Graph Metrics
7. 5% Registration & Hosting Data

## 7. Biography



**Vinit kumar gunjan** has completed his B.Tech degree in computer science from Trinity university in 2009. He is pursuing M.Tech in computer science from Sharda university, Greater Noida, Uttar Pradesh, India. His research interest includes Biological (DNA) development, Network & internet security Computing.

**Vinit kumar gunjan** has completed his B.Tech degree in computer science from Trinity university in 2009. He is pursuing M.Tech in computer science from Sharda university, Greater Noida, Uttar Pradesh, India. His research interest includes Biological (DNA) development, Network & internet security Computing.



**Pooja** has completed her B.Tech degree in IT from Chaudhary charan singh university in 2009. She is pursuing M.Tech in computer science from Sharda university, Greater Noida, Uttar Pradesh, India. Her research interest includes Web technology & Network security.



**Monika** has completed her M.Tech degree in computer science from KIIT university, Bhubaneswar in 2010. She is working as a Asst. Professor in Department of computer Sharda University, Greater Noida, Uttar Pradesh, India.

Pradesh, India. Her research interest includes Data Structures, Design and Analysis of Algorithms, Graph Theory, Discrete Mathematics, Optimization theory, Object Oriented programming & Artificial intelligence.



**Dr. Amit Kumar** is CEO and Chief Scientific Officer of BioAxis DNA Research Centre (BDRC) Pvt Ltd and Secretary, IEEE Hyderabad Section. Dr Amit Kumar is a member of IEEE, ISCB, APBIONET, IIIR and PRIB. He obtained his PhD in Applied Bioinformatics in 2007. Dr Kumar was nominated as “Pioneers in Genomic education 2010” by Ocimum Biosolutions Hyderabad and Gene Logic USA. He has organized, chaired and given invited talks in several National and International Conferences like PRIB 2007 Singapore, PRIB 2008 Australia, WCCI 2010 Barcelona Spain, DNA 2009 Andhra Pradesh Police academy and DNA 2010- Osmania University and several IEEE events etc.



**Dr Allam Appa Rao**, Vice-Chancellor of the JNTU Kakinada, is an iconic and towering personality in the field of education and research. His contributions to the field of Computer Engineering have been exemplary and spilled over into numerous other areas of science and technology, making him a pioneer of scientific advancements meant for the benefit of society. Dr Allam Appa Rao began his career in 1969 and went on to complete his Ph. D in Computer Engineering from Andhra University in the year 1984. This is the first Ph. D in Computer Engineering from Andhra University.

## 8. Acknowledgement

We are very thankful to the Research Technology Development Centre team Sharda University and the course coordinator for his ideas and excellent computing, Research & Development facilities at the University campus. In addition to Sharda University we also pay tribute to Dr IMS lamba for his moral support and the webmaster team of BioAxis DNA Research Centre, India who allowed us access to their website in order to implement our ideas.

## 9. References

- [1] (2011)Wikipedia[Online].Available:<http://en.wikipedia.org/wiki/PageRank>
- [2] S.Mukherjee, "A probabilistic model for optimal searching of the deep Web", 2003.
- [3] (2011)Wikipedia[Online]Available:<http://en.wikipedia.org/wiki/PageRank>
- [4] Zhen Liu and Philippe Nain, "Optimization issues in Web search Engines", IBM research,. 2006, VI, 981-1015, DOI: 10.1007/978-0-387-30165-5\_34
- [5] (2011) Affordable SEO services website. [Online]. Available: <http://www.affordable-seo-services.com/on-page-optimization.html>
- [6] (2011) Affordable SEO services website. [Online]. Available: <http://www.affordable-seo-services.com/off-page-optimization.html>
- [7] (2011) Search engine optimization website [Online] Available:[http://seo.yu-hu.com/glossary/off\\_page\\_optimization.htm](http://seo.yu-hu.com/glossary/off_page_optimization.htm)
- [8] (2011) Google website [Online] Available: <http://www.google.com/.../search-engine-optimization-Starter-guide.pdf>
- [9] Fuxue Wang; Yi Li; Yiwen Zhang; Coll. of Econ. & manage, "An empirical study on the search engine optimization technique and its outcomes," Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), Aug. 2011.
- [10] (2011) HubspotBlog [Online] Available: <http://blog.hubspot.com/blog/tabid/6307/bid/5014/Study-Shows-Business-Blogging-Leads-to-55-More-Website-Visitors.aspx>
- [11] (2011) Marketing tech blog [Online] Available:<http://www.marketingtechblog.com/how-to-optimize-a-press-release-for-search/>
- [12] (2011) Submit express website. [Online].Available: <http://www.submitexpress.com>
- [13] (2011)"Search Engine Ranking Factors V2". SEOmoz.org[Online].<http://www.seomoz.org/article/search-ranking-factors>.

# DSP Implementation of a Power Factor Correction Strategy for BLDC Motor Drive

R.Vijayarajeswaran  
Managing Director, Vi Microsystems Pvt. Ltd.  
Chennai 600096, India

## Abstract

The paper develops a power factor correction mechanism for a brushless dc permanent magnet (BLDC-PM) motor drive system through the use of a front end boost converter. It evolves a wave shaping mechanism to arrive at the sinusoidal nature for the input current in an effort to improve the input power factor. The theory is articulated using a closed loop algorithm to revolve around the operating range of the drive motor. The performance is evaluated on a MATLAB platform to elucidate the viability of the scheme in addition to highlighting its speed regulating capability. It steals the role of a Digital Signal Processor (DSP) to implement the proposed methodology and there from validate the results with a view to illustrate its practical applicability.

## 1. Introduction

Power Factor Correction (PFC) circuits appear to occupy the input stages in almost every medium and high power switching power supply systems operating at the voltage mains. The most often used configuration includes a boost converter at the primary stage, stabilizing the input to the second stage. Conventional off-line power converters with front - end diode-capacitor rectifier inherit a distorted input current waveform with high harmonic content. Though a variety of passive and active PFC techniques are in vogue, the passive techniques may be the best choice only in low-power and cost sensitive applications. Besides the dc voltage on the energy-storage capacitor in a single-stage PFC converter is not regulated.

The Brushless DC motors (BLDC) continues to attract the drive industry owing to its simplicity, low-cost and robust structure and it is suitable for variable-speed applications. The structure though simple draws a pulsating AC line current resulting in low power factor and high harmonic line current. However with the increasing demand for improved power quality, there is a definite need for better strategies to accomplish a high performance BLDC motor drive.

A novel power factor correction strategy suitable for brushless DC motors has been suggested [1]. It has been found to eliminate the use of boost unity power factor stage and bulk electrolytic capacitors. An algorithm for

power factor correction of direct torque controlled brushless DC motor drive in the constant torque region has been outlined [2]. An intelligent power factor correction methodology based artificial neural network has been proposed [3]. The dynamic characteristics of the brushless DC motor and the currents & voltages of inverter components have been analyzed through the use of fuzzy logic controller [4]. The current controlled mechanism has been found to allow the re-generative braking of BLDC motor and resultant improving the efficiency and lowering the acoustic noise [5]. It augurs the use of a power-factor-correction mechanism appropriately interfaced with a BLDC motor driver circuit to arrive at the desired quality of power.

### 1.1. Problem Definition

The primary focus is to design a control algorithm that envisage to improve the input power factor of an inverter fed BLDC motor drive in addition to regulating its speed. It attempts to incorporate an AC - DC boost converter at the front end and builds a comprehensive closed loop strategy to reshape the nature of the input current wave. The addition of boost interface adds to the advantages of high efficiency and power density.

## 2. Proposed Strategy

A boost converter is controlled by pre - calculated duty cycles to land at sinusoidal input current waveform. The input voltage feed - forward compensation enables the output voltage to be insensitive to the input voltage variation and guarantees the sinusoidal input current even if the input voltage is distorted. The methodology is evaluated through simulation and validated using a Digital Signal Processor (DSP) based prototype over the entire operating range.

The power module of the proposed approach is displayed in Fig.1 shows the boost converter cascaded with an inverter to power the BLDC motor. The Hall sensors



imbibed in the rotor provide the necessary feedback to regulate the speed.

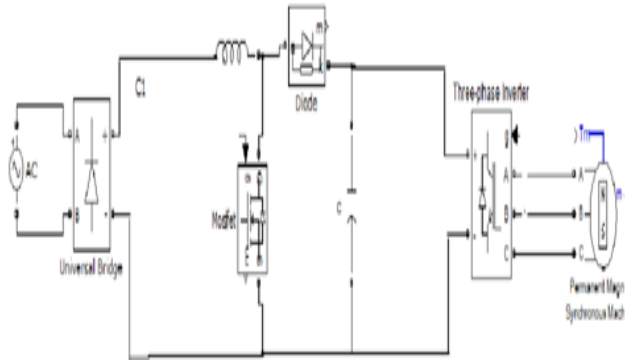


Fig. 1 Power Module

### 3. Control Algorithm

The objective of the control scheme of the boost converter is to regulate the power flow ensuring the tight output voltage regulation as well as unity input power factor. The modeling equations of various components of the converter system are formulated separately to develop a comprehensive model for their performance evaluation. The supply system under normal operating conditions can be modeled as a sinusoidal voltage source of amplitude  $V_m$  and frequency  $f_s$ . The instantaneous voltage is given as:

$$V_s(t) = V_m \sin \omega t \quad (1)$$

Where

$$\omega = 2\pi f_s t$$

And  $t$  is the instantaneous time.

A template  $u(t)$  is estimated for converter topologies with AC side inductor, from the sensed voltage

$$u(t) = v_s(t) / V_m \quad (2)$$

$u(t)$  for converter topologies with dc side inductor is obtained from:

$$u(t) = |V_s(t)| / V_m \quad (3)$$

The converters are modeled using first order non-linear differential equations. The number of equations is equal to the number of energy storage components in the system. The Single-phase boost PFC converter is modeled using two differential equations for inductor current  $i_L$  and DC link capacitor voltage  $V_{DC}$ .

$$p i_L = (v_d - v_p) / L - r(i_L / L) \quad (4)$$

$$p v_{dc} = (i_p - v_{dc} / R) / C_d \quad (5)$$

Where  $p$  is the differential operator ( $d/dt$ ),  $r$  is the internal resistance of the inductor  $L$ ,  $V_d$  is the rectified line voltage of diode rectifier output,  $R$  is the resistance of the load and  $V_p$  is the PWM voltage across the switch and is defined as

$$V_p = V_{DC} (1 - s) \quad (6)$$

$i_p$  is the current through the boost diode and is defined as

$$i_p = i_L (1 - S) \quad (7)$$

Where  $S$  is the switching signal obtained from current regulation loop. Its value is 1 (ON) or 0 (OFF) depending upon the state of the switch.

The Fig.2 shows the Schematic diagram of AC-DC boost type PF controller where the output of voltage regulator is limited to a safe value and forms the amplitude of input reference current. This reference amplitude is then multiplied to a template of input voltage to synchronize the reference with input voltage, as required for unity power factor operation. The inductor current is forced to track its reference current using current controller, which generates the appropriate gating signals for the active device.

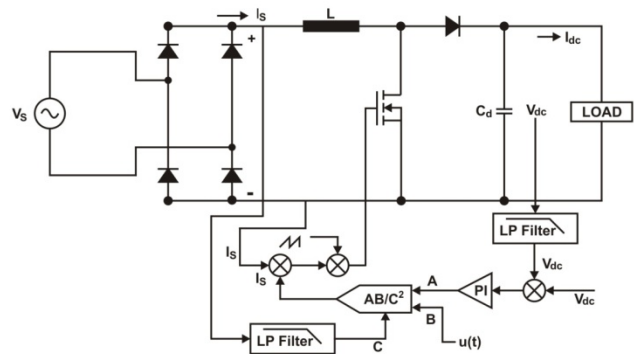


Fig. 2 Schematic diagram of AC-DC boost type PF controller

A proportional integral (PI) voltage controller is selected for voltage loop for tight regulation of the output voltage. The DC voltage  $V_{DC}$  is sensed and compared with set reference voltage  $V_{DC}^*$ .

The resulting voltage error  $v_{e(n)}$  at  $n$ th sampling instant is:

$$v_{e(n)} = V_{dc}^* - V_{dc}(n) \quad (8)$$

Output of PI voltage regulator  $v_{o(n)}$  at  $n$ th sampling instant is:

$$v_{0(n)} = v_{0(n-1)} + K_p(v_e(n) - v_e(n-1)) + K_i v_e(n) \quad (9)$$

Where  $K_p$  and  $K_i$  are the proportional and integral gain constants.  $v_e(n)$  is the error at the  $(n)$ <sup>th</sup> sampling instant. The output of the controller  $V_{0(n)}$  after limiting to a safe permissible value is taken as amplitude of reference supply current. The current regulation loop is required for active wave shaping of input current to achieve unity input power factor and reduced harmonics. The input voltage template  $B$  is obtained from the sensed supply voltage and is multiplied with the amplitude of reference source current  $A$  in the multiplier-divider circuit. Moreover, a component of input voltage feed forward  $C$  is also added to improve the dynamic response of the converter system to line disturbances. The resulting signal forms the reference for input current. The instantaneous value of the reference current is given as:

$$i_s^* = AB/C^2 \quad (10)$$

The inductor current error is the difference between the reference supply current and inductor current ( $i_{en} = i_s^* - i_s$ ). This error signal is amplified and compared to the fixed frequency carrier wave to generate the gating signals for power switches.

#### 4. Simulation

The scheme is simulated using MATLAB to investigate the performance of 310V, 4.2 A 3400 rpm and 1.1 HP BLDC motor. The Fig .3 depicts the input voltage and the input current at an operating point corresponding to 0.5 Kw, brought in phase due to the action of the PFC. The Comparisons of the input power factors over the entire range of load powers are elucidated through Fig. 4 to highlight the significant role of the PFC.

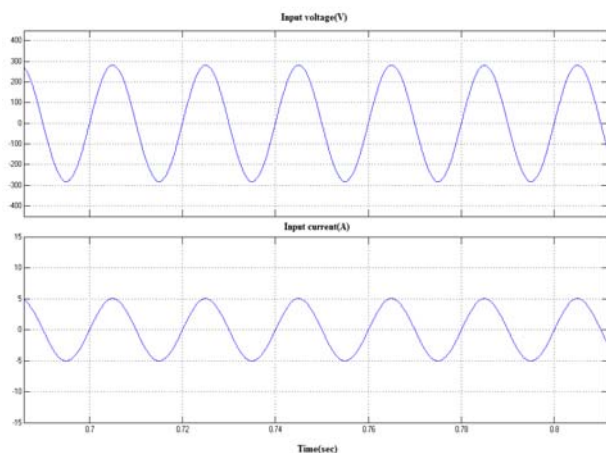


Fig .3 Input Voltage and current of PFC

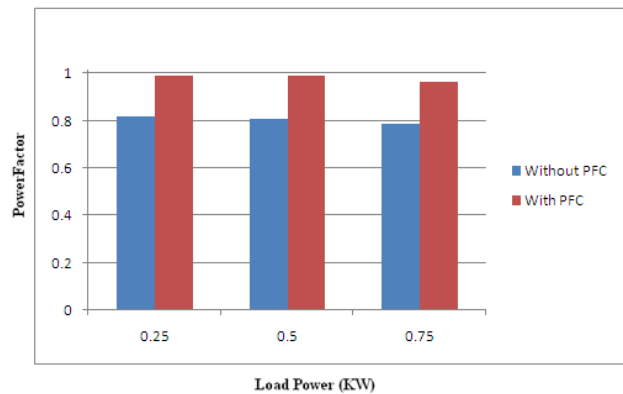


Fig .4 Comparison of Power Factor with and without PFC

#### 5. Hardware Implementation

The Performance of the algorithm is experimentally tested on the BLDC motor of similar ratings with DSP TMS320FC2812 based control to identify the numerous non-topological factors that affect the quality of current drawn by the converter. The power circuit is fabricated with IRFP460 MOSFET and MUR460 fast recovery diode with  $L_s = 1.1$  mH and  $C_d = 560$ mF. The converter is fed from AC lines through an autotransformer followed by an isolation transformer to provide variable input voltage and protection, respectively.

The DSPs are designed for closed loop control implementations are extensively used in areas of motor control, UPS, and motion control applications.

The Fig.5 shows a power factor corrector (PFC) stage interfaced to a TMS320LF2407A DSP. It is the AC-DC boost converter stage that converts the AC input voltage to a high voltage DC bus and maintains sinusoidal input current at high input power factor. As indicated in Fig.5, three signals (the rectified input voltage  $V_{in}$ , the inductor current  $I_{in}$ , and the DC bus capacitor voltage  $V_o$ ) are required to implement the control algorithm and the converter is controlled by two feedback loops. The average output DC voltage is regulated by a slow response 'outer loop' whereas, the inner loop shapes the input current is a much faster loop.

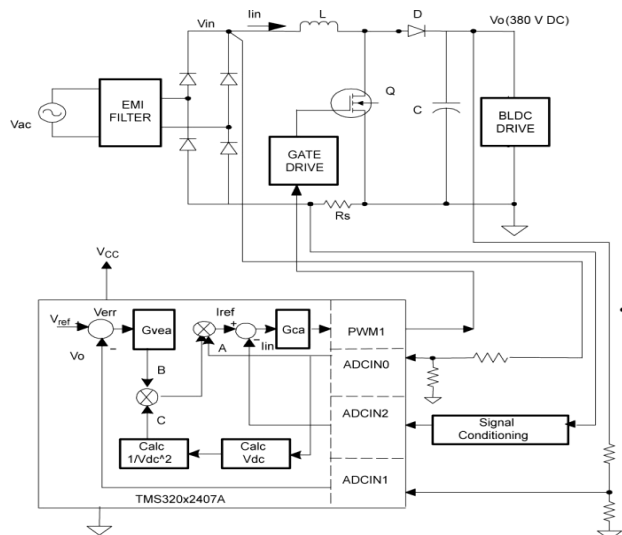


Fig. 5 TMS320LF2812 Controlled Power Factor Corrector (PFC) Stage for BLDC Drive

The instantaneous signals  $V_{in}$ ,  $V_o$  and  $I_{in}$  are all sensed and conditioned by the respective voltage and current sense circuits. The sensed signals are then fed back to the DSP via three ADC channels ADCIN0, ADCIN1, and ADCIN2 respectively. The rate at which these signals are sensed and converted by the ADC is called the control loop sampling frequency  $f_s$ . The digitalized sensed bus voltage  $V_o$  is compared to the desired reference bus voltage  $V_{ref}$ . The difference signal  $(V_{ref} - V_o)$  is then fed into the voltage loop controller  $G_{vea}$ . The digitized output of the controller  $G_{vea}$ , indicated as 'B', is multiplied by two other components, 'A' and 'C', to generate the reference current 'C' for the inner current loop. The component 'C' is calculated as,

$$C = \frac{1}{V_{DC} \times V_{DC}}$$

Where  $V_{dc}$  is the calculated average component of the sensed digitized signal  $V_{in}$ . In Fig.5,  $I_{ref}$  is the reference current command for the inner current loop.  $I_{ref}$  takes the shape of a rectified sine wave and its amplitude maintains the DC output voltage with a reference level  $V_{ref}$  against the variation in load and fluctuation in line voltage. The sensed digitized inductor current  $I_{in}$  is compared with the reference current  $I_{ref}$ . The difference between  $I_{ref}$  and  $I_{in}$  is passed into the current controller  $G_{ca}$ . The output of this controller is finally used to generate the PWM duty ratio command for the PFC switch.

The Fig.5 (a) shows the control loop block diagram of the DSP controlled PFC converter. In this figure, the voltage and current sense/conditioning circuits are replaced by their respective gain blocks.

These blocks are indicated as  $K_f$ ,  $K_s$  and  $K_d$ . The multiplier gain  $K_m$  is also added to the control block and it allows the adjustments of the reference signal  $I_{ref}$  based on the converter input operating voltage.

The inner loop is the current loop which is programmed by the reference current signal  $I_{ref}$ . The input to the current loop power stage is the duty ratio command  $d$  and its output is the inductor current  $I_{in}$ . The current controller  $G_{ca}$  is designed to generate the appropriate control output  $U_{ca}$  such that the inductor current  $I_{in}$  follows the reference current  $I_{ref}$ . The outer voltage loop is programmed by the reference voltage command  $V_{ref}$ . The input to the voltage loop power stage is  $U_{nv}$  (voltage controller output) and its output is the dc bus voltage  $V_o$ . The voltage controller  $G_{vea}$  is designed to generate the appropriate  $U_{nv}$  to control the amplitude of the reference current  $I_{ref}$  such that for the applied load current and line voltage, the bus voltage  $V_o$  is maintained at the reference level.

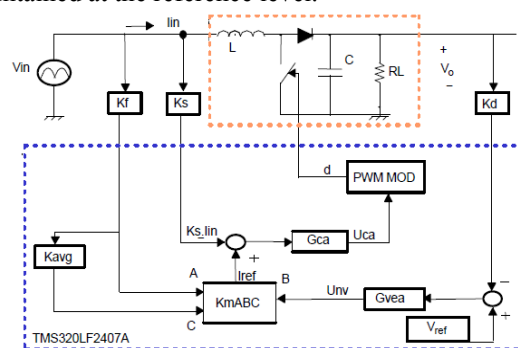


Fig. 5(a) Control Loop Block Diagram of the DSP Controlled PFC Stage

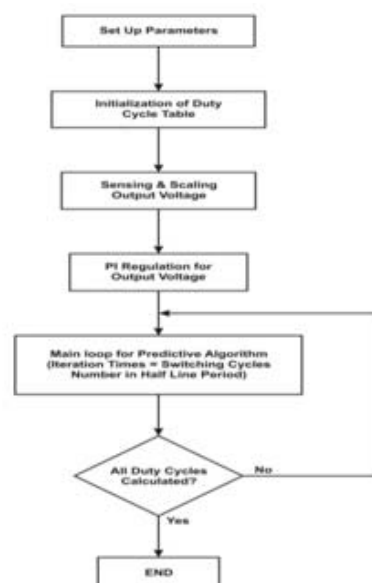


Fig. 6 Flow chart for DSP program

## 6. Experimental Results

The performance of the experimental prototype is investigated over the same operating range and the results obtained are displayed through Figs. 7 to 9. The PWM pulses to the power switches in the inverter at the chosen operating point are displayed in Fig. 7. The speed – time relationship established in Fig. 9 explains its regulatory action owing to the influence of the closed loop algorithm even when the BLDC motor is subjected to a sudden ten percent change both in reference speed and load. The input voltage and current waveforms in Fig.8 serves to validate the simulated response.

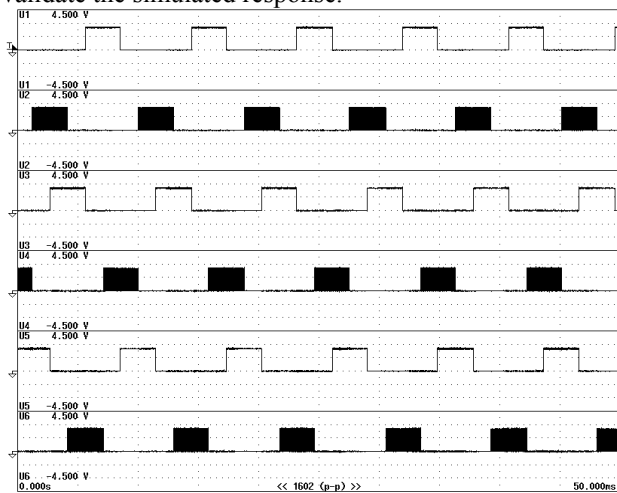


Fig .7 PWM pulses for inverter

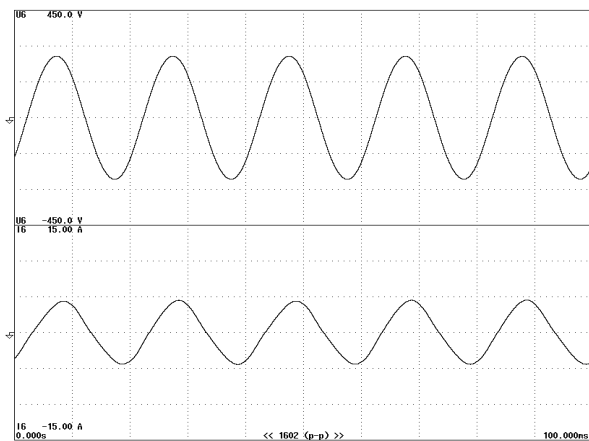


Fig .8 Source voltage and current wave

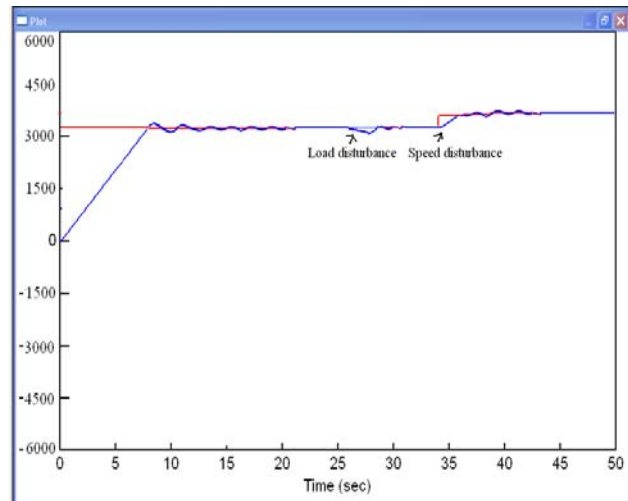


Fig .9 speed regulation graph

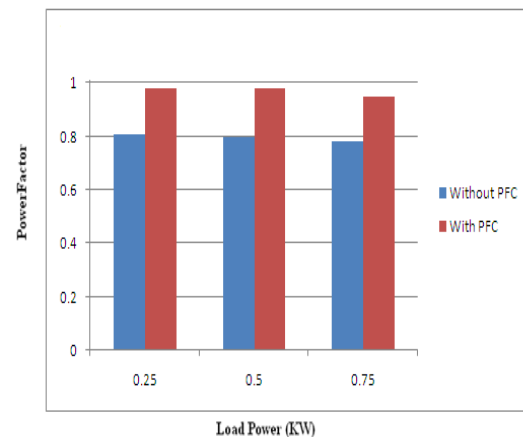


Fig .10 Comparison of Power Factor with and without PFC

The bar diagram displayed in the Fig.10 highlights the significant improvement in input powerfactor, adequately validated through DSP based prototype. With the load power allowed to vary across an appropriate operating range the experimental readings are compared with simulated results in Table.1 to validate the proposed methodology and highlight its suitability for practical applications.

S.no	Load power(KW)	Input current(A)	Input voltage(V)	Output current(A)	Speed(RPM)		Powerfactor	
					simulation	DSP	simulation	DSP
1	0.25	2.3	240	1.05	3400	3400	0.991	0.982
2	0.5	3.1	239	2.6	3400	3400	0.989	0.981
3	0.75	4.14	230	3.1	3400	3400	0.965	0.95

Table .1 Performance comparison in simulation and real time implementation

## Conclusion

An average current control algorithm has been developed for the BLDC drive to achieve input power factor through the use of a boost topology. The scheme has been formulated using high switching frequency PWM signals generated based on current feedback. The distortion in the input current has been corrected by triggering the power factor controller and shaping the input current wave into the desired sinusoid. A prototype has been constructed using a DSP controller board and ML4821 IC and the performance evaluated.

The experimental results have been found to adequately validate the simulated response and exhibit the suitability of the use of DSPs in this domain. The exercise has been found to illustrate the suitability of the proposed algorithm for practical systems and will go a long way in nurturing further innovative applications for BLDC motors.

## References

- [1]. Barkley, D. Michaud, E. Santi, A. Monti and D. Patterson , "Single Stage Brushless DC Motor Drive with High Input Power Factor for Single Phase Applications", 37<sup>th</sup> IEEE Power Electronics specialists conference, pp. 1-10, 2006.
- [2]. S.B.Ozturk and H.A .Oh Yang Toliyat, "Power Factor Correction of Direct Torque Controlled Brushless DC Motor Drive " , Industry Applications Conference, 42nd IAS Annual Meeting Conference Record of the IEEE, pp. 297 - 304 ,2007.
- [3]. R. Bayindir, S. Sagiroglu and I. Colak, "An intelligent power factor corrector for power system using artificial neural networks", Electric Power Systems Research, Vol 79, pp. 152-160, 2009.
- [4]. Mehmet Cunkas and Omer Aydođdu , "Realization of Fuzzy Logic Controlled Brushless DC Motor Drives using Matlab/Simulink", Mathematical and Computational Applications, Vol. 15, No. 2, pp. 218-229, 2010.
- [5]. G.MadhusudhanaRao and Dr. B.V.SankerRam, "speed control of BLDC motor with common current" , International Journal of Recent Trends in Engineering, Vol 2, No. 6, 2009.
- [6]. A.M. Tuckey, D.J. Patterson, "The design and development of a high power factor current source controller for small appliance brushless DC motors," Eleventh Annual Applied Power Electronics Conference and Exposition, 1996 (APEC '96), pp. 778 - 781, vol.2, 1996.
- [7]. E.Figueres, J.Benavent, M.Garcera, M.Pascual, "Robust control of power-factor-correction rectifiers with fast dynamic response," Industrial Electronics, IEEE Transactions on, pp. 66- 76, vol.52, no.1, 2005.
- [8]. Tan Chee Siong, Baharuddin Ismail, Siti Fatimah Siraj and Mohd Fayzul Mohammed, "Fuzzy Logic Controller for BLDC Permanent Magnet Motor Drives", International Journal of Electrical & Computer Sciences IJECS-IJENS,pp.13-18, Vol: 11 No: 02,2011.
- [9]. Atef Saleh Othman Al-Mashakbeh, "Proportional Integral and Derivative Control of Brushless DC Motor", European Journal of Scientific Research, pp.198-203 , Vol.35 , No.2 ,2009.
- [10]. Mehdi Nasri, Hossein Nezamabadi-pour, and Malihe Maghfoori," A PSO-Based Optimum Design of PID Controller for a Linear Brushless DC Motor", World Academy of Science, Engineering and Technology , pp. 211-215, 2007.
- [11]. M. Azizur Rahman , and Ping Zhou , "Analysis of Brushless Permanent magnet Synchronous Motors", IEEE Transactions On Industrial Electronics, pp 256-267 , VOL. 43, NO. 2, 1996.
- [12]. X.Li Q.Zhang and H.Xiao, "The design of brushless DC motor servo system based on wavelet ANN, "in Proc. Int. Conf. Machine Learning and Cybernetics, pp. 929-933, 2004.

**About Author:** Mr. R.Vijayarajeswaran, Chairman of the Vi Institute of technology doing his Phd., in Annamalai University, done is his M.E., from Government College of Engineering, Guindy. He started his career as Scientist at CEERI, Govt of India's Research Laboratory, Chennai. He has 30 years of rich experience in Evolving New Concepts, Design, Develop, Manufacture, Marketing, Finance Control and Administering of an Electronics related companies like M/S Pragathi Computer, Pondichery and M/S Vi Microsystems Pvt. Ltd., Chennai. He has Co-authored the book "**A Practical approach in DSP**", Published by "**New Age Publication**" and many of his papers are published in the National & international Journals and presented many papers in National & International Conference conducted by IEEE.

# Modified Secret Sharing over a Single Path in VoIP with Reliable Data Delivery

Mrs.K.Maheswari<sup>1</sup>, Dr.M.Punithavalli<sup>2</sup>

<sup>1</sup>Associate professor, Department of Computer Applications  
SNR SONS College, Coimbatore- 641006, Tamilnadu, India

<sup>2</sup>Director, Department of Computer Applications  
Sri Ramakrishna Engineering College, Coimbatore -641022  
Tamilnadu, India

## Abstract

Voice over Internet Protocol (VoIP) is a new fancy and up growing technology. A major change in telecommunication industry is VoIP. The transmission of Real time voice data is not as easy as ordinary text data. The real time voice transmission faces lot of difficulties. It suffers from packet loss, delay, quality and security. These factors will affects and degrade the performance and quality of a VoIP. This paper addresses the security and packet delivery ratio of a VoIP using modified secret sharing algorithm over a single path with reduced packet loss. The simulation results show that higher security and reduced packet loss is achieved in terms of end – to – end delay and packet delivery ratio. The user gets bad quality of VoIP at the receiver side. This makes the deployment of real time application a challenging task. To overcome these challenges in VoIP, several solutions have been reported already. The proper selection of active path in the routing protocol has a great impact in terms of packet delivery ratio and route discovery process. To provide end to end security between the source destination pair, the single path routing scheme is introduced.

**Keywords:** VoIP, Secret Sharing, packet loss, security, single path.

## 1. Introduction

Confidentiality is very important requirement for any kind of data transmission. The data in VoIP networks are not subject to eavesdropping. Preventing data from people who do not need to know. It is a packet switched and interactive network. The traditional Public Switched Telephone Network (PSTN) is circuit switched. The circuit switched network is secure one but the packet switched internet is not. It is designed with less security features. In conventional public switched telephone networks (PSTN), entire communication paths were administered by a few authorized telephone companies. It was therefore difficult for a malicious person to wiretap conversations over telephones because persons who were

allowed to access the network were carefully restricted. The recently grown internet protocol telephone or VoIP has multiple intermediates exist between the two endpoints (telephones). Therefore, the risk of man-in-the-middle attack increases. A message is divided into shares which are sending through a single path [Abdur Rashid Sang, et al.(2010)]. The modified shamir's secret sharing algorithm [A. Shamir (1979)] is implemented to provide reliable data delivery.

The transmission technology of VOIP must be in digital is shown in Figure 1. The caller's voice is digitized. The digitized voice is compressed and then separated into packets using complex algorithms [10]. These packets are addressed and sent across the network which is to be reassembled in the proper order at the destination. Again, this reassembly can be done by a carrier, and Internet Service Provider, or by PC.

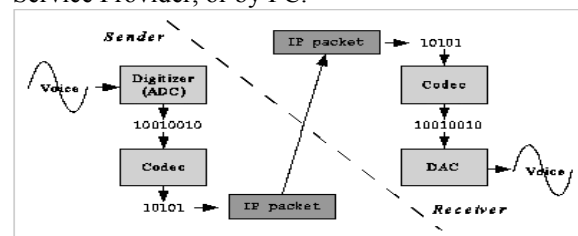


Figure 1 Transmission Technology of VoIP

During transmission on the Internet, packets may be lost or delayed, or errors may damage the packets. Conventional error correction techniques would request the retransmission of unusable or lost packets, but if the transmission is a real-time voice communication this technique obviously would not work, so sophisticated error detection and correction systems [Christos tachtatzis, et al (2008)] are used to create sound to fill in the gaps.

The fundamental idea of secret sharing is the secret message is sending through a single specified path using Ad-hoc On Demand distance Vector (AODV) routing [C.E.Perkins, et al (2001)] [M.K.Marina, et al

(2001)]. The enemy can easily compromise the message by troubling any one of the nodes all along the path. To solve this, the message is divided into shares or pieces. The pieces are sending through the specified path [Christos tachtatzis, et al (2008)].

A certain number of shares are used to reconstruct the original secret message. This is termed as Threshold secret sharing. Any shares less than threshold cannot do anything.

- Dividing the secret message into N multiple pieces called shares [Berry Schoenmakers (1999)][ Hanoch, et al (2006)]
- The enemy has to compromise at least T shares
- Designed for cheating detection and cheater identification
- Modified Shamir's Secret sharing scheme is implemented

The subjective performance [A.Baciocola, et al (2005)] of VoIP quality is predicted by E-model by an average listener combining the impairment caused by transmission parameters. The rating can be used to predict subjective user reactions, such as the Mean Opinion Score (MOS) [11].

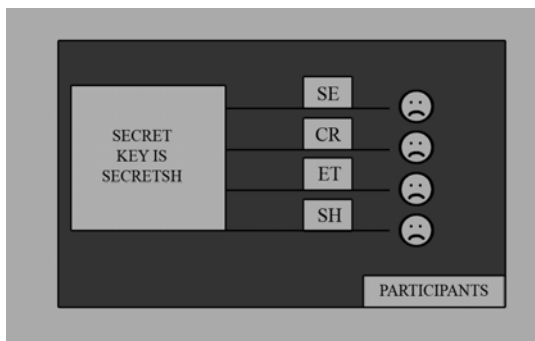


Figure 2 secret sharing in VoIP

According to ITU-T Recommendation, the E-model rating  $R$  is given by the following expression [ ITU-T Rec. G.107 (2003)]. $R = R_0 - I_s - I_d - I_e + A$ .

Where  $R$  -Transmission rating factor

$R_0$  - signal to noise ratio

$I_s$  - the combination of all impairments which occur more or less simultaneously with the voice signal

$I_d$  - the mouth-to-ear delay impairment factor

$I_e$  - equipment impairment factor

$A$  - The advantage factor or expectation factor

The resulting score is the transmission rating  $R$  factor, a scalar measure that ranges from 0 (poor) to 100 (excellent).  $R$  factor values below 60 are not

recommended [A. D. Clark (2001)] [R. G. Cole (2001)]. According to [ITU-T Rec. G.107 (2003)], the  $R$  factor is related to MOS as follows:

$$\begin{aligned} \text{For } R < 0 & \quad \text{MOS} = 1 \\ \text{For } 0 < R < 100 & \quad \text{MOS} = 1 + 0.035R + 7R(R-60)(100-R) \times 10^{-6} \\ \text{For } R > 100 & \quad \text{MOS} = 4.5 \end{aligned}$$

The E-Model not only takes in account the transmission statistics, but it also considers the voice application characteristics, like the codec quality, codec robustness against packet loss and the late packets discard. According to [A.Baciocola, et al (2005)], the above equations can be reduced to the following expression. where  $I_d$  is a function of the absolute one-way delay  $I_e$  is, in short, a function of the used codec type and the packet loss rate  $R = 93.4 - I_d - I_e$

Section 2 reviews the security threats of VoIP functions. This is followed by the threshold secret sharing scheme of VoIP in section 3. In section 4, the results are analyzed. Finally section 5 concludes the work.

## 2. Background Study

When the use of internet grows, automatically the complexity of the security problem increases. It becomes very difficult to solve the security problem. Actually, many application services do not consider the security. User authentication, confidentiality and integrity of signaling message or media stream are required for secure VoIP communication system.

The security threats are

- Eavesdropping and recording phone calls
- Tracking calls
- Stealing confidential information
- Modifying phone calls
- Making free phone calls
- Pranks / Practical jokes
- Board room bugging
- Sending spam (voice or email)
- Denial of service (DoS),
- Alteration of voice stream,
- Toll fraud,
- Redirection of call,
- Accounting data manipulation,
- Caller ID impersonation,
- Unwanted calls and messages

## 3. Threshold Secret Sharing Scheme in VoIP

This system divides a message into  $N$  pieces. Each  $N$  participant gets one share of the secret message

respectively. Any shares less than threshold cannot learn anything. The T (Threshold value of shares) out of N participants can rebuild the original secret message. This is called (T, N) threshold secret sharing scheme. The Shamir's Lagrange Interpolative Polynomial scheme is used to reconstruct the original. It is designed especially for identifying cheaters.

A secret sharing scheme consists of two algorithms

- Dealer
- Combiner

Dealer generates and distributes shares. The combiner collects and reconstructs the shares.

Shamir's construction for (T, N) secret sharing scheme is algebraic and is based on the polynomial interpolation. Assume K is the secret to be shared among N participants, S<sub>1</sub>, S<sub>2</sub>... S<sub>N</sub> are shares, P<sub>1</sub>, P<sub>2</sub>... P<sub>N</sub> can hold one share of the secret respectively. The dealer obtains the i<sup>th</sup> participant P<sub>i</sub>'s share S<sub>i</sub> by evaluating a polynomial of degree (T-1)

$$f(x) = K + a_1x + \dots + a_{T-1}x^{T-1} \pmod p \text{ at } x=I \text{ (} i=1,2,\dots,N\text{):}$$

$$P_i \rightarrow S_i = f(i) \tag{1}$$

Where

a<sub>1</sub>, a<sub>2</sub>, a<sub>3</sub>,...a<sub>T-1</sub> are coefficients which are selected randomly, part of a secret message

P is randomly chosen large prime number.

To indicate the security features of routes the vector P = [p<sub>1</sub>, p<sub>2</sub>, ..., p<sub>M</sub>] is used.

P<sub>i</sub> (i=1,2,...,M) is the probability that a route i is compromised. It is assumed that P<sub>1</sub> ≤ P<sub>2</sub> ≤ ... ≤ P<sub>M</sub>. The paths are ordered based on its cost value. The distribution of shares n = [n<sub>1</sub>, n<sub>2</sub>, ..., n<sub>M</sub>]

where

n<sub>i</sub> is the number of parts of a message sent through the route i. n<sub>i</sub> ≥ 0 and it is an integer.

$$\sum_{i=1}^M n_i = N \tag{2}$$

The probability that the message is compromised equal to the probability that T or more shares are seized.

The combiner side, the knowledge of minimum number of T shares, f(i<sub>1</sub>), f(i<sub>2</sub>), ..., f(i<sub>T</sub>), the original polynomial f(x) can be reconstructed by Lagrange interpolation.

$$f(x) = \sum_{j=1}^T S_{i_j} \cdot l_{i_j}(x) \pmod p \tag{3}$$

Where

$$l_{i_j}(x) = \prod_{k=1, k \neq j}^T \frac{x - i_k}{i_j - i_k} \tag{4}$$

At the source

$$T = \sum_{i=1}^N h(S_i) P^{2(i-1)} + \sum_{i=1}^{N-1} c p^{2i-1} \tag{5}$$

Where c is a positive constant, security features are added with shares.

At the destination

$$T^* = \sum_{i=1}^N h(s_i^*) p^{2(i-1)} \tag{6}$$

For each S<sub>i</sub>\*

$$\left[ \frac{T - T^*}{p^{2(i-1)}} \right] \pmod p = 0 \tag{7}$$

Choosing the most appropriate values of (T, N) and allocating them on to the paths is very important. (T, N) threshold secret sharing algorithm is applied to the message at source. If one node compromise data, all the shares traveling through the node would be compromised.

Reactive, demand – driven algorithm is AODV (Ad hoc on demand Distance vector routing). It discovers a route to a destination only when it sends a packet for forwarding to that destination. The discovered routes are maintained by route maintenance procedures.

A link has a limited life time. The link will expire when the two end nodes are in out of transmission range. In on-demand routing protocols link status will not be updated until they are used. The broken link will cause a number of route errors and generates a packet loss. Therefore, each link is given an appropriate life time. If this value is too small, link expires too soon. If the value is very large, links break early before the timers expire. It degrades the overall performance.

A predefined static life time is assigned for T1 seconds. In static life time scheme, the two clock time attributes are used.

- Born state
- Last used state

Born state indicates a new link is found in the route. The last used state indicates timestamp when the link is last used to forward a packet. There are two situations that will cause a link to be removed from the route.

- Route error is received or link is broken
- Timeout

If a link is removed because of the reception of a route error, the life time is calculated as



$$l = \text{CurrentTime}() - \text{link}[i,j].\text{born} \quad (8)$$

If it is removed because of timeout, the life time  $l$  is calculated as

$$l = \text{link}[i,j].\text{lastused} - \text{link}[i,j].\text{born} \quad (9)$$

LIFETIME is a variable indicating the estimation of the link lifetime. It is always assigned a static value. The modified algorithm is given below

**Step 1:** Create set  $S$ , which includes all the possible network security state vectors

$S = s_1, s_2, \dots, s_m$ . There should be totally  $2^M - 2$  elements in set  $S$ .

**Step 2:** Calculate  $Pstate(s)$  for each element  $s$  according to

$$Pstate(s) = \prod_{i=1}^M P_i^{s_i} (1-p)^{1-s_i} \quad (10)$$

Where  $i$  varies from  $1, 2, \dots, m$ .

Create set  $S2$ , which includes  $[1, 1, \dots, 1]$  only initially.

**Step 3:** Create set  $A$ , which include all the possible share allocation vectors

$$n = n_1, n_2, \dots, n_m$$

To reduce the size of  $A$

- $N \geq n_1 \geq n_2 \geq \dots \geq n_m \geq 0$
- $\sum n_i = N$  where  $i$  varies from  $1, 2, \dots, m$ .

**Step 4:** All the remaining elements in  $A$  are optimal share allocations if  $[1, 1, \dots, 1]$  is

the only element in set  $S2$ ; or they are sub-optimal share allocations if more elements present in set  $S2$ .

**Step 5:** Distributing the secrets in time domain basis by sending out the shares over a certain period of time. The link is estimated by its appropriate static lifetime value [in seconds].

**Step 6:** If the value assigned is very small, the link will expire too soon. At the same time if the value of lifetime is too big, there may be a route error. This will degrade the overall performance.

**Step 7:** Choosing the optimal value of static life time shows the performance of this algorithm

## 4. Results and Discussion

The important performance metric is End-to-End Delay. It is also called as Packet Latency. This is calculated by the time of packet sent at the sender and received at the receiver. This calculation is not only based on this but also the packets that are successfully delivered at the receiver without any loss of information.

When network traffic is very high, there may be a chance of packet latency at the receiver. So the success depends on the channel capacity. If the channel is more capable and error free, then there is no latency of packets. So the optimized lifetime value shows the result of low packet latency.

The results confirms that the small static lifetime value causes increasing number of route request and decreased number of route error. The Delay gets increased if the static lifetime is large and is shown in figure 3.

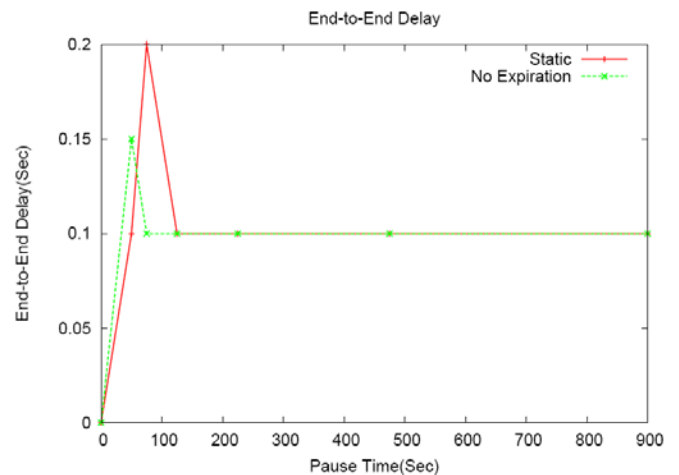


Figure 3 End – to – End Delay

The packet latency is calculated for packets that are successfully delivered. The transmission delay, propagation delay and queuing delay are the delay impairments that exist in IP networks. There are two types of latency.

- Protocol takes to discover a route to a destination
- Latency for a sender to recover when a route used breaks

It shows the average delay (time) in milliseconds spent to deliver each data packet.

$$\text{Average End-End Delay} = \text{TimeDelay} / \text{PacketReceived}$$

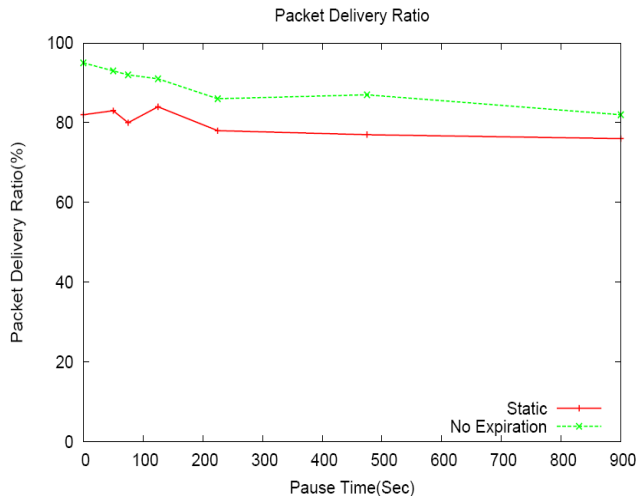


Figure 4 Packet Delivery Ratio

The application level performance metric is Packet Delivery Ratio (PDR). It is the ratio of packets that are received at the destination and sent at the source. It shows the ratio of total packets received at destination nodes, to total packets which are sent by source nodes.

$$PDR = \text{Packets received} / \text{Packets Sent} \times 100$$

The packets may be dropped due to route error. If there is no alternate path, the packet may be dropped. It shows the number of data packets which were dropped during their journey to destination. To reduce packet loss, a small lifetime value is favored. A small life time value reduces route error and increases route requests. Therefore more data is in transmission with less route error.

The other type of packet drop is due to heavy collisions. When the traffic is very high, the packet loss caused by collision becomes more rigorous. The Time versus Packet Delivery Ratio is shown in Figure 4. The performance of packet delivery ratio in high traffic significantly affects the routing overhead.

## 5. Conclusions

The streaming of audio or video content over the Internet is a challenging task. This is due to the fact that the Internet is a packet switched network with a little quality of service (QoS) guarantee. The major challenge of the VoIP network is maintaining quality as well as security. This work shows the result in a better performance. In unipath routing, only a single route is used between a source and the destination. The most commonly used protocols are Ad hoc On-Demand Distance Vector (AODV). The simulation results show that the reduction of packet loss and improvement of security. The performance is satisfied in terms of quality.

But the increased Delay and security are again a greater risk.

## References

- [1] Abdur Rashid Sang., Jianwei Liu., Zhiping Liu. performance comparison of single path and multipath routing protocol in MANET with selfish behavior, (2010).
- [2] A.Baciocola, C.Cicconetti, G.Stea. User-level performance evaluation of voip using ns2, (2005)
- [3] Berry Schoenmakers. A simple publicly verifiable secret sharing scheme and its application to electronic voting, CRYPTO 99 vol 1666 of lecture notes in Computer science, springer-verlag, pp: 148-164, (1999).
- [4] A. D. Clark. Modeling the Effects of Burst Packet Loss and Regency on Subjective Voice Quality, Columbia University IP, Telephony Workshop, (2001).
- [5] R. G. Cole and J. H. Rosenbluth. Voice over IP Performance Monitoring, ACM SIGCOMM, (2001) .
- [6] Christos tachtatzis, David harle. performance evaluation of multipath and single path routing protocols for mobile ad-hoc networks, (2008).
- [7] Hanoch, Levy, Haim, Zlatokrilov. the effect of packet dispersion on voice applications in IP networks, IEEE / ACM transactions on networking, vol.14, issue: 2, pp: 277-288, (2006) .
- [8] ITU-T Rec. G.107. The E-Model, A Computational Model for Use in Transmission Planning, (2003).
- [9] M.K.Marina, S.R.Das, performance of routing caching strategies in dynamic source routing, INt. Conf. on distributed computing system ICDCS, (2001) .
- [10] C.E.Perkins, E.M.Belding-Royer, S.R.Das, Ad-hoc On Demand distance Vector (AODV) routing, (2001) .
- [11] A. Shamir, "How to Share a Secret", Communications of the ACM, pp: 612- 613, (1979).



### First Author

K. Maheswari received her B.sc (Computer Science) from Madurai Kamaraj University and MCA. M.Phil. from Bharathidasan University. She is pursuing her Ph.d at Bharathiar University. She is currently working as an Associate Professor in the Department of Computer Applications, SNR Sons College, and Coimbatore. She has 16 years of teaching experience. She has presented research papers in several national and international conferences. She has published many research papers in various international journals. Her research interest is VoIP and network security.



### Second Author

Dr. M. Punithavalli received the Ph.d., degree in Computer Science from Alagappa University, Karaikudi in May

2007. She is currently serving as the Director of the Computer Applications Department, Sri Ramakrishna Engineering College, Coimbatore. Her research interest lies in the area of Data mining, Genetic Algorithms and Image Processing. She has published more than 10 Technical papers in International, National Journals and conferences. She is Board of studies member in various universities and colleges. She is also reviewer in International Journals. She has given many guest lectures and acted as chairperson in conference. Currently 10 students are doing Ph.D., under her supervision.

# Spoken Word Recognition Strategy for Tamil Language

AN. Sigappi<sup>1</sup> and S. Palanivel<sup>2</sup>

<sup>1</sup> Department of Computer Science and Engineering, Annamalai University  
Annamalainagar - 608 002, Tamilnadu, India

<sup>2</sup> Department of Computer Science and Engineering, Annamalai University  
Annamalainagar - 608 002, Tamilnadu, India

## Abstract

This paper outlines a strategy for recognizing a preferred vocabulary of words spoken in Tamil language. The basic philosophy is to extract the features using mel frequency cepstral coefficients (MFCC) from the spoken words that are used as representative features of the speech to create models that aid in recognition. The models chosen for the task are hidden Markov models (HMM) and autoassociative neural networks (AANN). The HMM is used to model the temporal nature of speech and the AANNs to capture the distribution of feature vectors in the feature space. The created models provide a way to investigate an unexplored speech recognition arena for the Tamil language. The performance of the strategy is evaluated for a number of test utterances through HMM and AANN and the results project the reliability of HMM for emerging applications in regional languages.

**Keywords:** *Speech recognition, Mel frequency cepstral coefficients, Hidden Markov models, Autoassociative neural networks.*

## 1. Introduction

Speech is the most common form of communication among human beings and it is their intuitive ability to recognize a spoken word, phrase, or sentence effortlessly. The benefits of spoken language interfaces are obvious due to the inherent natural communication approach present in them. However speech recognition systems for regional languages spoken in developing countries with rural background and low literacy rates appears to be still evolving.

Tamil is a classical Dravidian language that is in existence for several hundreds of years. It is classified as agglutinative language and spoken largely by people living in TamilNadu, parts of Srilanka, Singapore, and Malaysia. Tamil language imbibes twelve vowels and eighteen consonants, which combine to form two hundred and sixteen compound characters. In addition to this, there is a special letter called aaytham. Thus, a total of two

hundred and forty seven letters constitute the standard Tamil alphabet. The other significant feature present in Tamil language is the presence of unique liquid, which sounds 'zh'. The information required to perform the basic speech processing tasks is implicitly present in the speech. Owing to the fact that human beings are endowed with both speech production and perception mechanisms, the need for processing the speech signal does not arise. However there is a need to process the speech signal in light of the view that a machine is part of the communication chain.

## 2. Related Work

A grapheme-based automatic speech recognition system that jointly models phoneme and grapheme information using Kullback-leibler divergence based HMM has been presented and investigated for English language using DARPA Resource Management (RM) corpus [1]. A continuous speech recognizer using a group delay based two level segmentation algorithm has been developed to extract the accurate syllable units from the speech data [2]. Isolated style syllable models have been built for all unique syllables using samples from annotated speech in Tamil language. A formant tracking algorithm has been underlined using the phoneme information in the acoustic speech signal [3]. A robust coupled HMM-based audio video speech recognition (AVSR) system has been developed. The experimental results have been found to record a remarkable increase in the recognition rate compared to the only video based automatic speech recognition systems [4]. A prototype for speech based health information access by low literate community health workers has been developed. The experiences from a pilot study involving the use of community workers in a rural health center has been reported [5]. A privacy

preserving speech recognition model that serves to preserve the privacy between one party with private speech data and one party with private speech recognition models has been realized using HMM [6]. A host of methodologies for effective speech recognition have been articulated and evaluated using SPINE corpus. The use of parallel banks of speech recognizers have been found to improve the performance of recognition [7].

In spite of the urgent need for automation in all domains, the development of strategies for speech recognition in regional languages is still perceived to be cumbersome due to various issues such as non-availability of speech corpus for training purpose, complexity in the language, lack of phoneme recognizers and difficulty in creating a speech corpus with necessary transcriptions. Though this ordeal is in focus over a period of time, it still remains a challenge and efforts are required to accomplish this with greater precision and reliability. It is proposed to develop a strategy through which a spoken Tamil word can be recognized.

### 3. Problem Definition

It is an inert requirement to formulate a methodology through which computer systems can assimilate and recognize what a person speaks. A text and speaker dependent medium-sized vocabulary speech recognition mechanism is designed with an ability to recognize one hundred railway station names uttered in Tamil language. It echoes with it a focus to extricate its performance using HMM and AANN models and evolve a platform suitable to acclaim the spoken word from a chosen set of test samples.

### 4. Components of a Speech Recognition System

The constituents of a typical speech recognition system seen in Fig.1 include a feature extraction component and a template or statistical classifier. The main task of the feature extraction component is to extract features from a speech signal so as to represent the characteristics of the speech signal and yield a few numbers of coefficients that are grouped together to form a feature vector. Subsequent to feature extraction, the sequence of feature vectors is sent to a template or statistical classifier which selects the most likely sequence of word or phonemes.

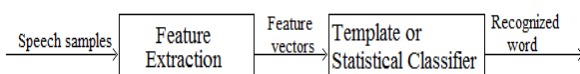


Fig. 1. Components of a speech recognition system

The fundamental issue in speech processing is the manner in which the specific features are extracted in order to perform the desired speech processing tasks. The human ear resolves frequencies non-linearly across the audio spectrum and it is thus desirable to obtain the nonlinear frequency resolution. Mel frequency cepstral coefficients (MFCC) appears to be one of the most successful feature representations in speech recognition related tasks, obtains the coefficients through a filter bank analysis. The mel-cepstrum exploits auditory principles, as well as the decorrelating property of the cepstrum [8]. Fig.2 illustrates the computation of MFCC features for a segment of speech signal [9]. The stages involved in the extraction of features are preemphasis, frame blocking, windowing, filter bank analysis, logarithmic compression, and discrete cosine transformation.

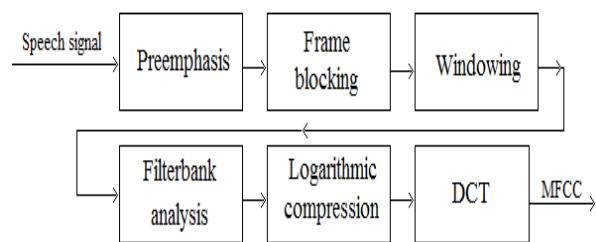


Fig. 2. Computation of MFCC features

**(i)Preemphasis:** Preemphasis is initiated to spectrally shape the signal so as to obtain similar amplitude for all formants. The speech signal is preemphasised by filtering the speech signal with a first order FIR filter whose transfer function in the  $z$ -domain is

$$H(z) = 1 - \alpha z^{-1}, 0 \leq \alpha \leq 1 \quad (1)$$

The preemphasised signal is related to the input signal in time domain using the relation. The preemphasis coefficient  $\alpha$  lies in the range  $0 \leq \alpha < 1$ .

$$\hat{s}(n) = s(n) - \alpha s(n-1) \quad (2)$$

**(ii)Frame blocking:** The statistical characteristics of a speech signal are invariant only within short time intervals of articulatory stability. The preemphasized signal is blocked into frames of  $N$  samples (frame size), with adjacent frames being separated by  $M$  samples (frame shift). If the  $l^{th}$  frame of speech is denoted by  $x_l(n)$  and there are  $L$  frames within the entire speech signal, then

$$x_l(n) = \hat{s}(Ml + n), \quad \begin{matrix} 0 \leq n \leq N-1 \\ 0 \leq l \leq L-1 \end{matrix} \quad (3)$$

**(iii) Windowing:** The next step is to window each frame so as to minimize the signal discontinuities at the beginning and end of the frame. The window is selected to taper the signal at the edges of each frame. If the window is defined as  $w(n)$ ,  $0 \leq n \leq N-1$ , then the result of windowing the signal is

$$x_1(n) = x(n)w(n), \quad 0 \leq n \leq N-1 \quad (4)$$

Hamming window is a good choice in speech recognition, considering that the subsequent operation in the feature extraction process integrates all the closest frequency lines. The Hamming window takes the form

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \leq n \leq N-1 \quad (5)$$

**(iv) Filter bank analysis:** A Fourier transform is obtained for each frame of the speech signal from which the magnitude is then weighted using a series of filter frequency responses. The center frequencies and bandwidths are chosen to roughly match those of the auditory critical band filters, that follow the mel scale, defined by

$$Mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \quad (6)$$

The filters are collectively called as a Mel scale filter bank and the frequency response of the filter banks simulate the perceptual processing performed within the ear.

**(v) Logarithmic compression:** The filter outputs obtained from filter bank analysis are compressed by a logarithmic function so as to model the perceived loudness of a given signal intensity.

$$X_{m(lm)} = \ln(X_{m}), \quad 1 \leq m \leq M \quad (7)$$

where  $X_{m(lm)}$  is the logarithmically compressed output of the  $m^{th}$  filter.

**(vi) DCT:** Discrete cosine transform (DCT) is thereafter applied to the filter outputs and the first few coefficients are grouped together as a feature vector of a particular speech frame. If  $p$  is the order of the mel scale cepstrum, the feature vector is obtained by considering the first  $p$  DCT coefficients. The  $k^{th}$  MFCC coefficient in the range  $1 \leq k \leq p$  can be expressed as

$$MFCC_k = \sqrt{\frac{2}{M}} \sum X_{m(lm)} \cos(\pi k(m - 0.5)M) \quad (8)$$

## 5. Hidden Markov Models (HMM)

Hidden Markov models (HMMs) are widely used in automatic speech recognition applications because of their accuracy in recognition. A Hidden Markov model is characterized by the following [10]:

(i)  $N$ , the number of hidden states in the model. The individual states are indicated as  $S = S_1, S_2, \dots, S_N$ , and the state at time  $t$  as  $q_t$ .

(ii)  $M$ , the number of distinct observation symbols per state. The observation symbols are denoted as  $V = v_1, v_2, \dots, v_M$ .

(iii) The state transition probability distribution  $A = a_{ij}$ , where

$$a_{ij} = P[q_{t+1} = S_j | q_t = S_i], \quad \begin{matrix} 1 \leq i \\ j \leq N \end{matrix} \quad (9)$$

(iv) The observation probability distribution in state  $j$ ,  $B = b_j(k)$ , where

$$b_j(k) = P[v_k \text{ at } t | q_t = S_j], \quad \begin{matrix} 1 \leq j \leq N, \\ 1 \leq k \leq M \end{matrix} \quad (10)$$

(v) The initial state distribution  $\pi = \{\pi_i\}$  where

$$\pi_i = P[q_1 = S_i], \quad 1 \leq i \leq N \quad (11)$$

A complete specification of a HMM is as follows:

$$\lambda = \{a_{i,j}, b_{j,k}, \pi\}, \quad \begin{matrix} \sum_j a_{i,j} = 1, \forall i \\ \sum_k b_{j,k} = 1, \forall j \end{matrix} \quad (12)$$

An isolated word recognizer using HMM is shown in Fig. 3. A training set of  $K$  occurrences and the features extracted from each occurrence of the word constitutes an observation sequence for every word in the vocabulary. The HMM constructed for each word estimates the model parameters  $(A, B, \pi)$  that optimises the likelihood of the training set observation vectors. The observation sequence for the test utterance is determined from the speech signal from where the most likelihood calculation is made for all possible models, using which the word whose model likelihood appears to be the highest is selected.

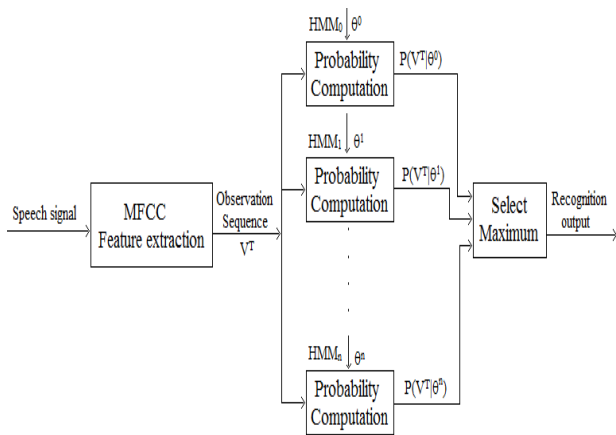


Fig. 3. Isolated word Recognizer using HMM

### 6. Autoassociative Neural Networks (AANN)

Autoassociative neural network models (AANNs) are feed forward neural networks bestowed with an ability to perform identity mapping of the input space [11] and are increasingly used in speech processing applications. The distribution capturing capability of the AANN model is described using the five layer AANN model as shown in Fig. 4. This model comprises of three hidden layers in which the processing units in the first and third hidden layer are nonlinear whereas the units in the second hidden layer can be either linear or nonlinear. The first layer in the network is the input layer, the third is the compression layer and the last is the output layer. The second and fourth layers contain more units than the input layer, while the third layer is with fewer units than the first or fifth layer. In any case, the number of units in the input and output layers are the same.

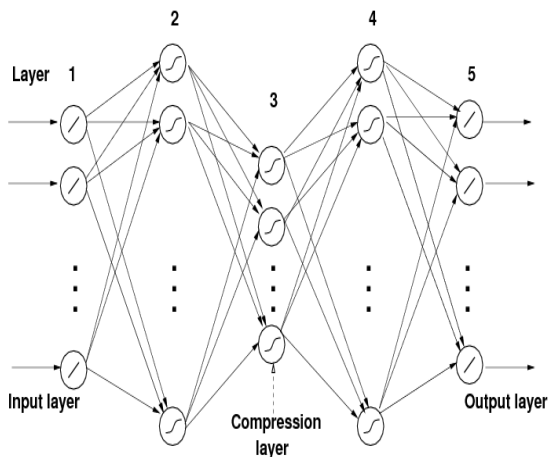
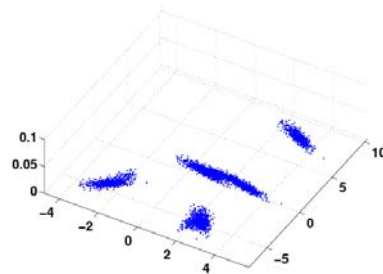


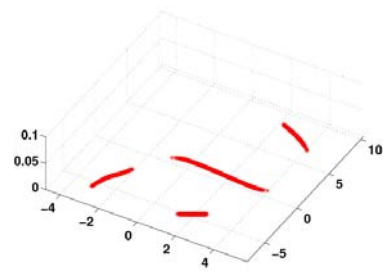
Fig. 4. A five layer autoassociative neural network model

The cluster of points in the input space determines the

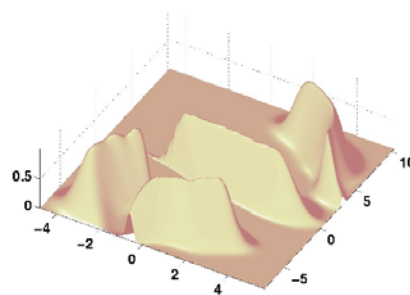
shape of the hypersurface obtained by the projection onto the lower dimension space, as the error between the actual and the desired output vectors is minimized. The space spanned by the one dimensional compression layer in Fig.5(b) corresponds to the two dimensional data shown in Fig.5(a) and the network structure  $2L\ 10N\ 1N\ 10N\ 2L$ , where  $L$  denotes a linear unit and  $N$  denotes a non linear unit. The integer values indicate the number of units present in that layer. The nonlinear output function for each unit is  $\tanh(s)$ , where  $s$  is the activation value of the unit. The network is trained using the backpropagation algorithm. The AANN thus captures the distribution of the input data depending on the constraints imposed by the structure of the network.



(a)



(b)



(c)

Fig. 5 Distribution capturing capability of AANN. (a) Artificial two dimensional data. (b) Two dimensional output of the AANN model. (c) Probability surfaces realized by the AANN.

The error for each input data point is plotted in the form of some probability surface as given in Fig. 5(c). The error  $e_i$  for the data point  $i$  in the input space is plotted as  $p_i = \exp(-e_i/\alpha)$ , where  $\alpha$  is a constant. Though  $p_i$  is not strictly a probability density function, the resulting surface is termed as probability surface. The plot of the probability surface shows a larger amplitude for a smaller error  $e_i$ , indicating a better match of the network for that data point. The constraints imposed by the network is seen by the shape the error surface takes in both cases. An ideal expectation pertaining to the distribution of data is oriented to achieve the best probability surface, best defined in terms of some measure corresponding to a low average error.

## 7. Experimental Results

### 7.1 Speech Corpus Creation

The strategy is evolved for a speech recognition task, with a view to identify the spoken utterances of specific words in Tamil language. The vocabulary includes one hundred railway station names in TamilNadu. The procedure necessitates the creation of an in-house speech corpus that contains the audio recordings of the words uttered by speakers. An omnidirectional microphone is used for recording the voices and the recording is carried out in a laboratory environment. Wavesurfer software is used for the recording purpose and the recorded files are stored in .wav format. The sampling rate is chosen to be 16 KHz. The speakers are chosen to be in different ages ranging from twenty one to sixty and of both genders in order to ensure good training and replicate the actual scenario. Twenty speakers are chosen and a speaker is made to utter each station name five times and hence a total of ten thousand samples constitute the speech corpus. The experiments are conducted using the database created, from which eight thousand samples are used for training purpose and the remaining two thousand samples for testing the performance of the methodology.

### 7.2 Feature Representation

The input speech data is preemphasised using a first order digital filter, with a preemphasis coefficient of 0.97. It is then segmented into ten millisecond frames with an overlap of fifty percent between adjacent frames and windowed using Hamming window. The twelve MFCC coefficients are thereafter extracted from the filter bank output through the use of DCT. In addition to the log energy, the 0'th cepstral parameter  $C_0$  is appended to give thirteen MFCC coefficients. The time derivatives are

added to the basic static parameters thus obtained, in order to further enhance the performance of the speech recognition strategy. The thirteen first order regression coefficients, known as delta coefficients and the thirteen second order regression coefficients, referred to as acceleration coefficients are appended to the static coefficients to yield thirty nine mel frequency cepstral coefficients. The problem of separating the background silence from the input speech is another factor that accords a significant impact on the implementation of a speech recognition system [10]. It is typically done on the basis of signal energy and signal durations. A frame is judged to be nonsilent if its total energy is less than 40 dB below the maximum total energy computed across all the frames in the utterance. The other frames are considered to be silent frames and are not taken into account for purpose of model creation.

### 7.3 Model Construction

The 39 dimensional MFCC feature vectors extracted from the nonsilence frames of the speech signal corresponding to each word are given as input to estimate the parameters of HMM and is implemented using HTK toolkit. The HTK tools HInit and HRest provide isolated word style training using a flat start mechanism and a HMM for each station name is generated individually. Once an initial set of models are created, the tool HRest performs a Baum-Welch reestimation of the entire training set. Each of the models are reestimated until no change occurs in the state transition probabilities and finally the required models are made available to represent each station name. The structure of the AANN model used in the experiments is chosen by systematically varying the number of units in the second and third layers and by varying the number of epochs required for training the network.

### 7.4 Performance

The performance of the strategy presented in this work is evaluated through the voice samples recorded from the speakers that are set aside for testing. The recognition rate is used as the performance measure and it is defined as the number of words correctly recognized. It is given by the equation

$$r = \frac{c}{t} \times 100 \quad (13)$$

where  $r$  represents the recognition rate,  $c$  the number of words correctly recognized during testing, and  $t$  the total number of words in the vocabulary.

The experiment using HMM is conducted by varying the



number of states in the model and also the number of mixtures in each state. The results shown in Fig.6 indicate that the HMM with 5 states and 4 mixtures in each state yields a recognition rate of 95.0%. Similarly the performance of the strategy using AANN is evaluated by varying the number of units in the second ( $N_s$ ) and third ( $N_t$ ) layers. These results seen in Fig.7 explain that the network structure  $39L\ 80N\ 30N\ 80N\ 39L$  trained for 600 epochs offers the best results with a recognition rate of 90.0%.

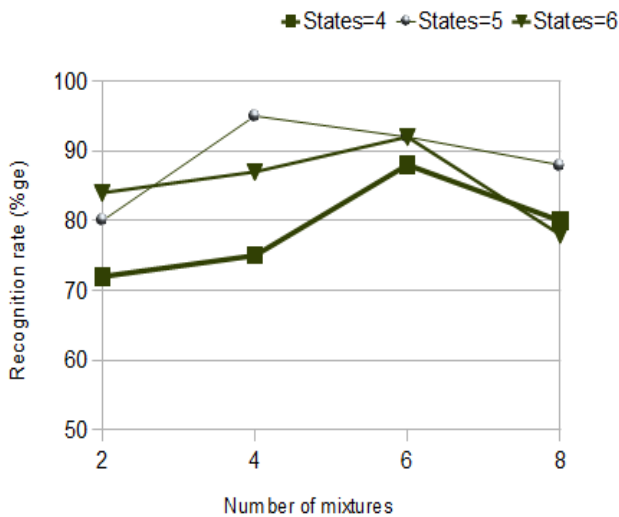


Fig. 6. Recognition results for various states and mixtures in the HMM

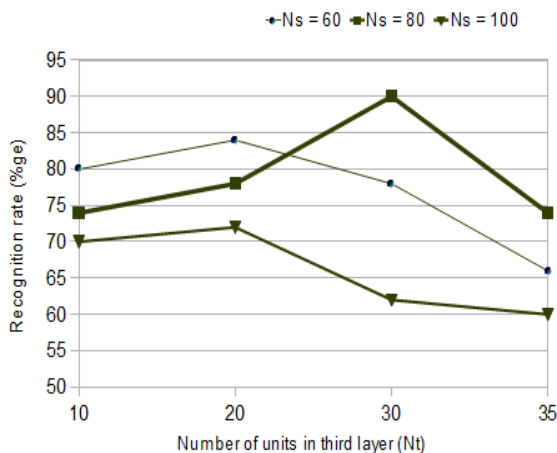


Fig. 7. Performance of different AANN network structures

The results summarized in Fig.8 indicate that HMM is a reliable model promising for speech recognition applications in comparison to AANN. It also follows that the responses are closely related to the strength of the samples used. It precisely points to the need for a rich corpus to accomplish the worthy use of the structure.

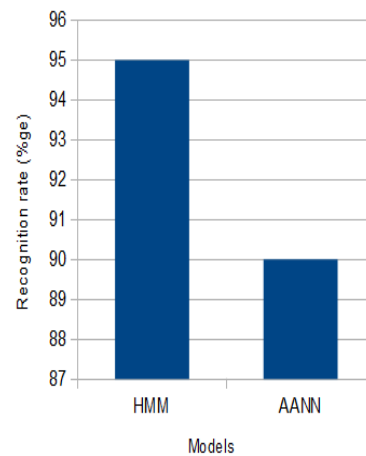


Fig. 8. Comparison of HMM and AANN results

#### 4. Conclusion and Future Work

A strategy for recognizing spoken words in Tamil language has been developed. The strategy has been evolved using HMM and AANN models constructed using the mel frequency cepstral coefficient features obtained from the speech signal. The HMM with 5 states and 4 mixtures has been found to yield better recognition results in comparison to the AANN model with a structure  $39L\ 80N\ 30N\ 80N\ 39L$ . The performance of the strategy has been found to explicitly portray the suitability of HMM over AANN for a speech recognition task in Tamil language. The consistency and robustness of the approach have been proved to be its highlights and allow its use in unforeseen environments. Besides if new features can be explored to characterize the speech signal more accurately it will go a long way in arriving at higher recognition rates.

#### References

- [1] Mathew Magimai.-Doss, Ramya Rasipuram, Guillermo Aradilla, and Herve Bourlard, "Grapheme-based automatic speech recognition using KL-HMM", in Proceedings of Interspeech, Aug 2011.
- [2] A. Lakshmi, and Hema A. Murthy, "A syllable based continuous speech recognizer for tamil", in Intl. Conf. on Spoken Language Processing, Sept 2006.
- [3] M. Lee, J. van Santen, B. Mobius, and J. Olive, "Formant tracking using context-dependent phonemic information", IEEE Transactions on Speech and Audio Processing, Vol. 13, No. 5, Sept 2005, pp. 741-751.
- [4] Yashwanth .H., Harish Mahendrakar, and Suman David, "Automatic speech recognition using audio visual cues", in IEEE India Annual Conferenc, Dec 2004, pp. 166-169.

- [5] Jahanzeb Sherwani, Nosheen Ali, Sarwat Mirza, Anjum Fatma, Yousuf Memon, Mehtab Karim, Rahul Tongia, and Roni Rosenfeld, "HealthLine: Speech-based access to health information by low-literate users", in IEEE/ACM Intl. Conference on Information and Communication Technologies and Development, Dec 2007.
- [6] Paris Smaragdis and Madhusudana Shasashanka, "A framework for secure speech recognition", IEEE transactions on Audio, Speech, and Language Processing, Vol.15, No.4, May 2007, pp.1404-1413.
- [7] John H. L. Hansen, Ruhi Sarikaya, Umit Yapanel, and Bryan Pellom, "Robust speech recognition in noise: An evaluation using SPINE corpus", in EUROSPEECH 2001, Sept 2001, pp. 4148-4153.
- [8] S. B. Davis, and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol.28, No.4, Aug. 1980, pp.357-366.
- [9] "The HTK Book", Cambridge University Engineering Department, 2002.
- [10] Lawrence R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition", Proceedings of the IEEE, Vol.77, N0.2, Feb 1989, pp.257-285.
- [11] B. Yegnanarayana, and S. P. Kishore, "AANN: an alternative to GMM for pattern recognition", Neural Networks, Vol.15, No.3, April 2002, pp.459-469.

**AN. Sigappi** received her Bachelors degree in Computer Science and Engineering from Annamalai University in 1993 and her Masters degree in the same discipline from Anna University in 2001. She is currently pursuing her doctoral research work at Annamalai University. Her career includes both administrative and academic experience spanning over 15 years and she is presently working as Associate Professor in the Department of Computer Science and Engineering at Annamalai University. Her research interest includes speech and image processing, software engineering, management information systems and intelligent systems.

**S. Palanivel** received the B.E(Hons) degree in Computer Science and Engineering from Bharathidasan University in 1989 and followed it up with Masters degree in the same discipline from Bharathiar University in 1994. He completed his Ph.D in Computer Science and Engineering from the Indian Institute of Technology Madras in the year 2005. He is currently serving as Associate Professor in Computer Science and Engineering at Annamalai University. He carries with him 17 years of teaching experience and over 20 publications in international conferences and journals. His research interests include speech processing, image and video processing, pattern classification and neural networks.

# Implementation of Genetic Algorithm in Predicting Diabetes

S.Sapna<sup>1</sup>, Dr.A.Tamilarasi<sup>2</sup> and M.Pravin Kumar<sup>3</sup>

<sup>1</sup> Assistant Professor,  
Department of Master of Computer Applications,  
K.S.R College of Engineering, Tiruchengode-637215  
Tamilnadu, India

<sup>2</sup> Prof. & Head, Department of Master of Computer Applications,  
Kongu Engineering College, Perundurai,  
Tamilnadu, India

<sup>3</sup> Assistant Professor,  
Department of Electronics and Communication Engineering,  
K.S.R College of Engineering, Tiruchengode-637215  
Tamilnadu, India

## Abstract

Data Mining aims at discovering knowledge out of data and presenting it in a form that is easily compressible to humans. Data Mining represents a process developed to examine large amounts of data routinely collected. The term also refers to a collection of tools used to perform the process. One of the useful applications in the field of medicine is the incurable chronic disease diabetes. Data Mining algorithm is used for testing the accuracy in predicting diabetic status. Fuzzy Systems are been used for solving a wide range of problems in different application domain Genetic Algorithm for designing. Fuzzy systems allows in introducing the learning and adaptation capabilities. Neural Networks are efficiently used for learning membership functions. Diabetes occurs throughout the world, but Type 2 is more common in the most developed countries. The greater increase in prevalence is however expected in Asia and Africa where most patients will likely be found by 2030.

**Keywords:** Data Mining, Diabetes, Fuzzy Systems, Genetic Algorithm (GA), Neural Networks.

## 1. Introduction

Health and Commonwealth Government have identified diabetes to be a significant and growing global public health problem with the expected incidence in Australia to increase from 4% to 10% by 2010. An estimated 40 million Indians suffer from diabetes, and the problem seems to be growing at an alarming rate. By 2020, the number is expected to

double and reach epidemic proportions, even as half the numbers of diabetics in India remain undiagnosed. Diabetes has debilitating consequences on many of the body's vital organs if remained unchecked and controlled, the biggest problem being that of eyesight. It affects eyes, kidney, heart and every single vital organ of the body. India has the dubious distinction of being the diabetic capital of the world. Home to around 33 million people with diabetes, 19% of the world's diabetic population is from India. Nearly 12.5% of Indian's urban populations have diabetes. The number is expected to escalate to an alarming 80 million by the year 2030. Amongst the chronic diabetic complications, diabetic foot is the most devastating result. Over 50,000 leg amputations take place every year due to diabetes in India. Diabetes patients can often experience loss of sensation in their feet. Even the smallest injury can cause infection that can be various serious. 15% of patients with diabetes will develop foot ulcers due to nerve damage and reduced blood flow. Diabetes slowly steals the person's vision. It is the cause for common blindness and cataracts. Cardiovascular diseases are rising. Nearly 3.8 crore cases were detected in 2005 and experts believe the number will go upto 6.4 crore by 2015.

Fuzzy Systems is used for solving a wide range of problems in different application domains. The use of Genetic Algorithms for designing Fuzzy Systems

allows us to introduce the learning and adaptation capabilities. The topic has attracted considerable attention in the Computation Intelligence community. The paper briefly reviews the classical models and the most recent trends for Genetic Fuzzy Systems. Accurate and reliable decision making in oncological prognosis can help in the planning of suitable surgery and therapy, and generally, improve patient management through the different stages of the disease. To indicate that the reliable prognostic marker model than the statistical and artificial neural-network-based methods.

Genetic Algorithms (GAs) are considered as a global search approach for optimization problems. Through the proper evaluation strategy, the best "chromosome" can be found from the numerous genetic combinations. Although the GA operations do provide the opportunity to find the optimum solution, they may fail in some cases, especially when the length of a chromosome is very long. In this paper, a data mining-based GA is presented to efficiently improve the Traditional GA (TGA). By analyzing support and confidence parameters, the important genes, called DNA, can be obtained. By adopting DNA extraction, it is possible that TGA will avoid stranding on a local optimum solution. Furthermore, the new GA operation, *DNA implantation*, was developed for providing potentially high quality genetic combinations to improve the performance of TGA. Experimental results in the area of digital watermarking show that our data mining based GA successfully reduces the number of evolutionary iterations needed to find a solution.

Real-life data mining applications are interesting because they often present a different set of problems for data miners. One such real-life application that we have done is on the diabetic patients databases. In this paper, knowledge discovery on this diabetic patient database is discussed. A semi-automatic means for cleaning the diabetic patient database, and present a step-by-step approach to help the health doctors explore their data and to understand the discovered rules better. Generally in Asia about 47 percent of the population is diabetic. This disease has many side effects such as higher risk of eye disease, higher risk of kidney failure, and other complications. However, early detection of the

disease and proper care management can make a difference. To combat this disease a regular screening program for the diabetic patients. Patient information, clinical symptoms, eye-disease diagnosis and treatments are captured into a database. This leads naturally to the application of knowledge discovery and data mining techniques to discover interesting patterns that exist in the data. The objective is to find rules that can be used by the medical doctors to improve their daily tasks, that is, to understand more about the diabetic disease.

## 2. Fuzzy Systems

Application of fuzzy sets theory was recognized in the field of medicine, the uncertainty found in the process of diagnosis of disease that has most frequently been the focus of applications of fuzzy set theory. The desire to better understand and teach this difficult and important process of medical diagnosis has prompted attempts to model it with the use of fuzzy sets, These models vary in the degree to which they attempt to deal with different complicating aspects of medical diagnosis such as the relative importance of symptoms, the varied symptom patterns of different disease stages, relations between diseases themselves, and the stages of hypothesis formation, preliminary diagnosis, and final diagnosis within the diagnostic process itself. These models also form the basis for computerized medical expert systems, which are usually designed to aid the physician in the diagnosis of some specified category of diseases.

## 3. Genetic Algorithms

Genetic algorithm (GA) refers to a model introduced and investigated by John Holland in 1975 for adaptation processes of nature. Generally stated, a GA is any population based model that uses selection and recombination operators to generate new sample points in a search space. GA computationally utilizes a natural evolutionary process similar to the process first described by Charles Darwin in his "The Origin of Species", to solve a given problem. GA is a global search procedure that searches from one population of points to another. GA is a probabilistic search procedure, which is being frequently applied to difficult optimization and learning problems. There are two versions of the GA, namely the natural GA and the computational GA.

Genetic algorithms were inspired by the processes observed in natural evolution. They attempt to mimic these processes and utilize them for solving a wide

range of optimization problems. In general, genetic algorithms perform directed random searches through a given set of alternatives with respect to the given criteria of goodness. These criteria are required to be expressed in terms of an objective function, which is usually referred to as a fitness function.

Genetic algorithms require that the set of alternatives to be searched through be finite. If we want to apply them to an optimization problem where this requirement is not satisfied, the set involved and select an appropriate finite subset. It is further required that the alternatives be coded in strings of some specific finite length which consist of symbols from some finite alphabet. These strings are called chromosomes, the symbols that form them are called genes, and their set is called a gene pool. Genetic algorithms search for the best alternative in the sense of a given fitness function through chromosomes evolution. Basic steps in genetic algorithms are shown in figure 1.

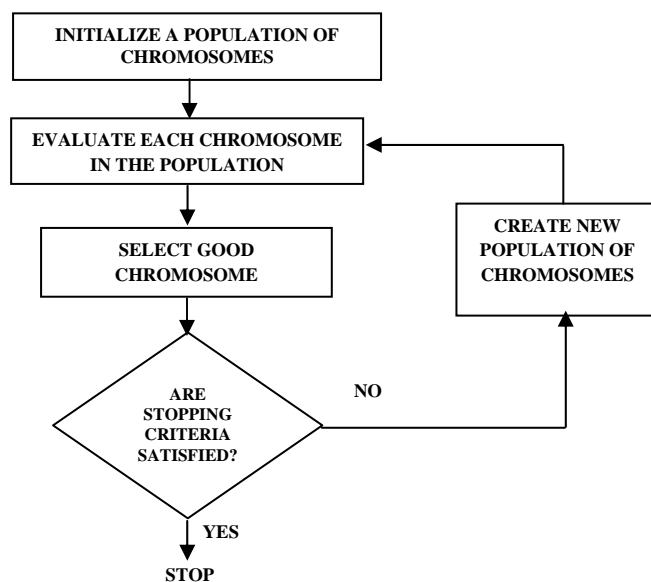


Fig 1. Flowchart of Genetic Algorithms

Genetic Algorithms search for the best alternative (in the sense of a given fitness function) through chromosomes' evolution. Basic steps in genetic algorithms figure. First, an initial population of chromosomes is randomly selected. Then each of the chromosomes in the population is evaluated in terms of its fitness (expressed by the fitness function). Next, a new population of chromosomes is selected from the given population by giving a greater change to select chromosomes with the high fitness. This is called *natural selection*.

The new population may contain duplicates. If given stopping criteria (e.g., no change in the old and new population, specified computing time, etc.,) are not met, some specific, genetic – like operations are performed on chromosomes of the new population. These operations produce new chromosomes, called offspring's. The same steps of this process, evaluation and natural selection, are then applied to chromosomes of the resulting population. The whole process is repeated until given stopping criteria are met. The solution is expressed by the best chromosome in the final population.

There are many variations on these basic ideas of genetic algorithms. To describe a particular type a genetic algorithm in greater detail, let  $G$  denote the gene pool, and let  $n$  denote the length of strings of genes that form chromosome. That is, chromosomes are  $n$  - tuples in  $G^n$ . The size of the population of chromosomes is usually kept constant during the execution of genetic algorithm. That is, when new members are added to the population, the corresponding the number of old members are excluded. Let  $m$  denote this constant population size. Since each population may contain duplicates of chromosomes, we express populations by  $m$ -tuples whose elements are  $n$ -tuples from the set  $G^n$ . Finally, let  $f$  denote the fitness function employed in the algorithm.

3.1 Genetic algorithm which is iterative, consists of the following six steps

1. Select an initial population,  $P^{(k)}$ , of a given size  $m$ , where  $k=1$ . This selection is made randomly from the set  $G^n$ . The choice of value  $m$  is important. If it is too large, the algorithm does not differ much from an exhaustive search; it is too small, the algorithm may not reach the optimal solution.
2. Evaluate each chromosome in population  $P^{(k)}$  in terms of its fitness. This is done by determining for each chromosome  $x$  in the population the value of the fitness function,  $f(x)$ .
3. Generate a new population  $P_n^{(k)}$ , from the given population  $P^{(k)}$  by some procedure of natural selection. We describe only one possible procedure of natural selection, which is referred to as deterministic sampling. According to this procedure, we calculate the value  $e(x) = mg(x)$  for each  $x$  in  $P^{(k)}$ , where  $g(x)$  is a relative fitness defined by the formula.

$$g(k) = \frac{f(x)}{\sum_{n \in P^{(k)}} f(x)}$$

Then the number of copies of each chromosome  $x$  in  $P^{(k)}$ , that is chosen for  $P_n^{(k)}$ , is given by the integer part of  $e(x)$ . If the total number of chromosomes chosen in this way is smaller than  $m$  (the usual case), then we select the remaining chromosomes for  $P_n^{(k)}$  by the fractional parts of  $e(x)$ , from the highest values down. In general, the purpose of this procedure is to eliminate chromosomes with low fitness and duplicate those with high fitness.

4. If stopping criteria are not met, go to step 5, otherwise stop.
5. Produce a population of new chromosomes  $P^{(k+1)}$ , by operating on chromosomes in population  $P_n^{(k)}$ . Operations that are involved in this step attempt to mimic genetic operations observed in biological systems. They include some or all the following four operations:

**A. Simple Crossover**

Given two chromosomes

$$x = (x_1, x_2, \dots, x_n), y = (y_1, y_2, \dots, y_n)$$

and an integer  $i \in N_{n-1}$ , which is called a crossover position, the operation of simple crossover applied to  $x$  and  $y$  replaces these chromosomes with their offspring's,

$$x' = (x_1, \dots, x_i, y_{i+1}, \dots, y_n),$$

$$y' = (y_1, \dots, y_i, x_{i+1}, \dots, x_n)$$

chromosomes  $x$  and  $y$ , to which this operation is applied, are called *mates*.

**B. Double Crossover**

Given the same chromosomes mates  $x, y$  as in the simple crossover and two crossover positions  $i, j \in N_{n-1} (i < j)$ , the operation of double crossover applied to  $x$  and  $y$  replaces these chromosomes with their offspring's,

$$x' = (x_1, \dots, x_i, y_{i+1}, \dots, y_j, x_{j+1}, \dots, x_n),$$

$$y' = (y_1, \dots, y_i, x_{i+1}, \dots, x_j, y_{j+1}, \dots, y_n)$$

**C. Mutation**

Given a chromosome  $x = (x_1, x_2, \dots, x_n)$  and an integer  $i \in N_n$ , which is called a mutation position, the operation of mutation replaces  $x$  with

$$x' = (x_1, \dots, x_{i-1}, z, x_{i+1}, \dots, x_n)$$

where  $z$  is a randomly chosen gene from the gene pool  $G$ .

**D. Inversion**

Given a chromosome  $x = (x_1, x_2, \dots, x_n)$  and two integers  $i, j \in N_{n-1} (i < j)$ , which are called inversion positions, the operation of inversion replaces  $x$  with

$$x' = (x_1, \dots, x_i, x_j, x_{j-1}, x_{i+1}, x_{j+1}, \dots, x_n)$$

- E. Replace population  $P_n^{(k)}$  with  $P^{(k+1)}$  produced in Step 4, increase  $k$  by one, and go to Step 2.

**3.2 Sample Problem**

(a)  $k = 1$  : Step 2 and 3

Chromosome in $P^{(1)}$	Integers	Fitness	$g(x)$	$4g(x)$	Number of Selected Copies
00010	2	3.75	0.068	0.272	0
01001	9	12.94	0.292	1.168	1
10011	19	15.44	0.350	1.400	2
11000	24	12.00	0.291	1.164	1

(b)  $k = 1$  : Step 5

Chromosome in $P_n^{(1)}$	Mate (randomly Selected)	Crossover Site (randomly Selected)	Resulting Chromosomes in $P^{(2)}$
01001	10011	3	01011
10011	01001	3	10001
10011	11000	1	11000
11000	10011	1	10011

Similarly for the values of  $k = 2, 3$ , values are set to calculate the fitness value, mate and the crossover site.

(c)  $k = 2$  : Step 2 and 3

Chromosome in $P^{(2)}$	Integers	Fitness	$g(x)$	$4g(x)$	Number of Selected Copies
01011	11	14.44	0.250	0.100	0
10001	17	15.94	0.276	1.104	2
11000	24	12.00	0.207	0.828	1
10011	19	15.44	0.267	1.068	1

(d)  $k = 2$  : Step 5

Chromosome in $P_n(2)$	Mate (randomly Selected)	Crossover Site (randomly Selected)	Resulting Chromosomes in $P(3)$
10001	3	2	10000
10001	4	3	10011
11000	1	2	11001
10011	2	3	10001

(e)  $k = 3$ : Step 2 and 3

Chromosome in $P^{(3)}$	Integers	Fitness	$g(x)$	$4g(x)$	Number of Selected Copies
10000	16	16.00	0.274	1.096	1
10011	19	15.44	0.265	1.060	1
11001	25	10.94	0.188	0.752	1
10001	17	15.94	0.273	1.092	1

A crossover operation is employed in virtually all types of genetic algorithms, but the operations of mutation and inversion are sometimes omitted. Their role is to produce new chromosomes not on the basis of the fitness function, but for the purpose of avoiding a local minimum. This role is similar to the role of a disturbance employed in neural networks. If these operations are employed they are usually chosen with small probabilities. The mates in the crossover operations and the crossover positions in the algorithm are selected randomly. When the algorithm terminates, the chromosome in  $P^{(k)}$  with the highest fitness represents the solution.

### 3.3 Natural Genetic Algorithm

The natural genetic algorithm is as follows :

- randomly generate an initial population  $M(0)$
  - loop
- a. Compute and save the fitness  $u(m)$  for each individual  $m$  in current population  $M(t)$ .
  - b. Define the selection probabilities  $p(m)$  for each individual  $m$  in  $M(t)$  (so that  $p(m)$  is proportional to  $u(m)$ ).
  - c. Generate  $M(k+1)$  by probabilistically selecting individuals from  $M(t)$  to produce a new population via genetic operators.

Fuzzy Genetic Algorithm can be implemented to check the patients affected by diabetes based upon the fitness value and the accuracy chromosome value.

### 3.4 Diabetes

Most of the food we eat is converted to glucose, or sugar which is used for energy. The pancreas secretes insulin which carries glucose into the cells of our bodies, which in turn produces energy for the perfect functioning of the body. When you have diabetes,

your body either doesn't make enough insulin or can not use its own insulin as well as it should. This causes sugar to build up in your blood leading to complications like heart disease, stroke, and neuropathy, poor circulation leading to loss of limbs, blindness, kidney failure, nerve damage, and death.

#### 3.4.1 General Symptoms of Diabetes

- Increased thirst
- Increased urination - Weight loss
- Increased appetite - Fatigue
- Nausea and/or vomiting - Blurred vision
- Slow-healing infections - Impotence in men

#### 3.4.2 Types of Diabetes

Type I - Diabetes also called as *Insulin Dependent Diabetes Mellitus (IDDM)*, or *Juvenile Onset Diabetes Mellitus* is commonly seen in children and young adults however, older patients do present with this form of diabetes on occasion. In type 1 diabetes, the pancreas undergoes an autoimmune attack by the body itself therefore; pancreas does not produce the hormone insulin. The body does not properly metabolize food resulting in high blood sugar (glucose) and the patient must rely on insulin shots. Type I disorder appears in people younger than 35, usually from the ages 10 to 16.

Type II - Diabetes is also called as *Non-Insulin Dependent Diabetes Mellitus (NIDDM)*, or *Adult Onset Diabetes Mellitus*. Patients produce adequate insulin but the body cannot make use of it as there is a lack of sensitivity to insulin by the cells of the body. Type II disorder occurs mostly after the 40.

Gestational Diabetes- Diabetes can occur temporarily during *Pregnancy* called as *Gestational Diabetes* which is due to the hormonal changes and usually begins in the fifth or sixth month of pregnancy (between the 24th and 28th weeks). Gestational diabetes usually resolves once the baby is born. However, 25-50% of women with gestational diabetes will eventually develop diabetes later in life, especially in those who require insulin during pregnancy and those who are overweight after their delivery.

#### 3.4.3 Diagnostic Tests

- Urine Test
- Fasting Blood Glucose Level
- Post Prandial Blood Sugar
- Random Blood Glucose Level
- Oral Glucose Tolerance Test
- Glycosylated Hemoglobin (HbA1c)

## 4. Computation Genetic Algorithm

The major difference between the natural and computational GA is that at some point in the loop, termination conditions are checked and the process is terminated if a termination condition is reached. Genetic Algorithm (simplified): Step of the algorithm

- a) Initialize the population
- b) Calculate the population's fitness
- c) While the number of generation is not the maximum number of generations do:
  - a) Select all the solutions whose genetic material will propagate to the next generation
  - b) Perform the crossover operation
  - c) Perform the mutation operation
  - d) Calculate the new population's fitness
  - e) Get population statistics

A genetic algorithm begins with a collection of solutions to the problem being solved, called a population. Along with each individual in the population is an associated fitness value, or measure of solution quality. The larger the fitness value, the better the solution. With the initial population generated and the fitness calculated, the algorithm iterates through a series of five operations which progressively improves the quality of the solutions in the population. These steps are:

### 1. Selection

The algorithm spins a "roulette wheel" to randomly select two individuals from the population whose genetic material will propagate to the next generation. This roulette wheel is not a fair wheel - solutions with better fitness's are more likely to be chosen.

### 2. Crossover (recombination operators)

The pairs of solutions chosen through selection, certain randomly chosen pairs undergo crossover, i.e., have their genetic material combined to create a new pair of solutions. Each of the newly created solutions inherits characteristics from both parents and is placed in the new population. Those solutions not chosen to be combined through crossover are simply copied into the new population. As its name implies, a crossover operator forms new chromosomes by combining (generally) two chromosomes with a (usually) predetermined crossover probability,  $p_c$ .  $p_c$  depends on the problem and other parameters, but it is often taken about 70-80%. Crossover is the main search operator of GAs.

### 3. Mutation

After a new generation has been created through copy and crossover, a certain number of the new solutions are randomly chosen to experience mutation. This operation perturbs the gene pool by introducing new solutions not directly related to existing solutions through crossover.

### 4. Fitness

Next, all solutions in the new population have their fitness's calculated.

### 5. Population Statistics

The new population with its fitness values is evaluated to determine and record the best solution found to date during the execution of the genetic algorithm. One complete pass through all five of these steps is referred to as a generation and results in the creation of a new population of solutions, equal in size to the starting population. Once this is complete the new population is used as the starting point for the next iteration and the process is repeated. The computation concludes when either a certain number of generations has been completed or when a given solution quality has been reached. The new generated population will be equal in size to the starting population. The new population is used as the starting point for the next iteration and the entire process is repeated until some determined termination conditions are reached. These termination conditions are when a certain number of generations is generated or when a given solution quality is reached.

## 4.2 Generic Genetic Algorithm

Procedure GA\_IPD\_Run

Initialize\_Population ( $P_{old}$ )

// fills the chromosome of population  $P_{old}$  with 0's and 1's randomly. while termination criteria not satisfied do for each chromosome  $c_i$  in  $P_{old}$  do

Evaluate ( $c_i, P_{old}$ ) // runs chromosome  $c_i$  against every member of  $P_{old}$  includes itself to compute fitness end

Generate\_New\_Population ( $P_{new}, P_{old}$ )

// generate new population using

$P_{old} P_{old} \rightarrow P_{new}$  end, end

## 4.3 Algorithm for Generating New Population

Procedure Generate \_ New \_ Population ( $P_{old}, P_{New}$ )

$P_{New} \rightarrow 0$

while Size ( $P_{New}$ ) < Size ( $P_{old}$ ) do

// Selection

$c_1 \leftarrow \text{Select} (P_{old})$   $c_2 \leftarrow \text{Select} (P_{old})$

// Crossover

if  $P_c < r(\cdot)$  then // return random nos. in the interval (0,1)

//  $P_c$ : Crossover Probability Crossover ( $c_1, c_2$ ) // implements uniform crossover end // Mutation

for  $i = 1$  to chromosome\_length do

if  $r(\cdot) < P_m$  then //  $P_m$  Mutation Probability

// Chromosome swapping each bit at the corresponding position with fixed probability usually 0.5 percent

$c_{li} \leftarrow c_{li}$  // ith bit of the  $i$ th chromosome

end



if  $r(.) < P_m$  then  $c_{2i} \leftarrow c_{2i}$  end, end

$P_{New} \rightarrow P_{New} \triangleleft c_1 \triangleleft c_2 // \triangleleft$  . Inserts the chromosome on the right hand side to the population to the left hand side.

## 5. Conclusion

Inorder to obtain the accuracy of chromosome and to evaluate the diabetes in diabetic patient GA is implemented. The connection between fuzzy systems and genetic algorithms is bidirectional. In one direction, genetic algorithms are utilized to deal with various optimization problems involving fuzzy systems. One important problem for which genetic algorithms have proven very useful is the problem of optimizing fuzzy inference rules in fuzzy controllers. In the other direction classical genetic algorithms can be fuzzified. The resulting fuzzy genetic algorithms tend to be more efficient and more suitable for some applications. Research on complex diseases only seems to be approaching the final goal, the prevention and cure of the diseases, very slowly. Diabetes is a disease in which the body does not produce or properly use insulin. Insulin is a hormone that is needed to convert sugar, starches and other food into energy needed for daily life. The cause of diabetes continuous to be ambiguous although both genetics and environmental factors such as obesity and lack of exercise. Symptoms of low blood sugar, side effects, science of complication are to be noted else it leads to severe problems. Using GA optimization of chromosome is obtained and based on the rate of old population diabetes can be restricted in new population to get chromosomal accuracy.

## Reference

- [1] Diabetes and you your guide to living well with diabetes, Novo Nordisk, LEAD GROUP.
- [2] How to cut out all Diabetic Problems by 50% - The Alphabet way, Dr.Vinod Patel, Department of Diabetes and Endocrinology George Eliot Hospital, UK.
- [3] Genetic mapping of complex traits: the case of Type 1 diabetes, Päivi Onkamo, Diabetes and Genetic Epidemiology Unit, Department of Epidemiology and Health Promotion, National Public Health Institute and Division of Biometry, Rolf evanlinna Institute and Finnish Genome Center Faculty of Science University of Helsinki Academic, 2002.
- [4] Genetic Fuzzy Systems: Status, Critical Considerations and Future Directions, Francisco Herrera, International Journal of Computational Intelligence Research. ISSN 0973-1873 Vol.1, No.1 (2005), pp.59-67.
- [5] A Fuzzy Logic Based-Method for Prognostic Decision Making in Breast and Prostate Cancers, Huseyin Seker, *Student Member, IEEE*, Michael O.

Odetayo, Dobrila Petrovic, and Raouf N. G. Naguib, *Senior Member, IEEE*

- [6] TRANSACTIONS on Information Technology in Biomedicine, Vol. 7, No. 2, June 2003.
- [7] A Data Mining Based Genetic Algorithm, YI-TA WU1, YOO JUNG AN2, JAMES GELLER2 AND YIH-TYNG WU3, 2006 IEEE.
- [8] Data Mining Diabetic Databases: Are Rough Sets a Useful Addition?, Joseph L. Breault, MD, MPH, MS, Department of Health Systems Management, Tulane University, Department of Family Practice, Alton Ochsner Medical Foundation, joebreault@tulanealumni.net.
- [9] Parallel Medical Image Analysis for Diabetic Diagnosis, Yueh-Min Huang, E-mail: huang@mail.ncku.edu.tw, Shu-Chen Cheng, ROC, E-mail: kittyc@mail.stut.edu.tw, *Int. J. Computer Applications in Technology, Vol. 22, No. 1, 2005.*
- [10] Feature Subset Selection Using a Genetic Algorithm, Jihoon Yang and Vasant Honavar, Iowa state University, IEEE Intelligent Systems Exploration Mining in Diabetic Patients Databases : Findings and Conclusions, Wynne Hsu Mong Li Lee Bing Liu Tok Wang Ling, School of Computing, National University of Singapore, Lower Kent Ridge Road, Singapore 119260, {whsu, leeml, liub, lingtw} @comp.nus.edu.sg.

## Biography



**S.Sapna** received her B.Sc Degree, M.C.A and M.Phil Degree, Bharathiar University. She is currently working as Assistant Professor in the Department of MCA, K.S.R. College of Engineering. She has presented more than 20 papers on various topics including national, international conference and journals. She is a research scholar of Mother Teresa Women's University, Kodaikanal. Her research interest includes Soft computing, Data Mining, Mathematical Computations and Networks. She is a member of life member of ISTE and CSI.



**Dr.A.TAMILARASI**, currently serving as Prof. & Head, Department of M.C.A, Kongu Engineering College. She has published various papers in the filed of fuzzy logics. She has published various books in mathematical field. She is guiding several research scholars in her area of interest like Data Mining, Soft Computing and Networks.



**M.Pravin Kumar** received his B.E & M.E Degree, from Anna University. He is currently working as Assistant Professor in the Department of ECE, K.S.R. College of Engineering. He has presented more than 5 papers on various topics including national, international conference and journals. He is a research scholar of Anna University, Coimbatore. His research interest includes Soft Computing and Networks. He is a member of life member of ISTE.

# Influence of Side Effect of EBG Structures on the Far-Field Pattern of Patch Antennas

F.BENIKHLEF, N. BOUKLI-HACENE

Telecommunications Laboratory, Technologies Faculty, Abou-Bekr Belkaïd University  
Tlemcen, 13000, Algeria

Telecommunications Laboratory, Technologies Faculty, Abou-Bekr Belkaïd University  
Tlemcen, 13000, Algeria

## Abstract

The electromagnetic band gap structure always used as a part of antenna structure in order to improve the performance of the antenna especially for improves the gain and radiation pattern. In this paper, microstrip antenna is used due to the advantages such as easy and cheap fabrication, light weight, low cost, easy to feed, and better isolations among array elements, by suppressing surface wave modes. The two dominating side effects are the parasitic loading effect and cavity effect. The first causes the multi resonances antenna resulting in large bandwidth, the second effect is due to reflecting energy from EBG toward antenna and so decreasing the bandwidth. The EBG structure parameters and number of EBG rows is related to these effects.

In this paper, we propose a rectangular microstrip patch antenna with EBG substrate of different structure EBG parameters and number of EBG rows; we compare the performance of the proposed antenna with a conventional patch antenna, in a same parametric analysis with HFSS simulator.

**Keywords:** Patch antenna, surface wave, EBG structure, gain and bandwidth.

## 1. Introduction

With the drastic demand of wireless communication system and their miniaturization, antenna design becomes more challenging. Recently microstrip patch antennas have been widely used. In spite of its several advantages, they suffer from drawbacks such as narrow bandwidth; low gain and excitation of surface waves [1], to overcome these limitations of microstrip patch antennas two techniques have been used to suppress surface wave propagation, namely micromachining [2] and periodic structures called the electromagnetic bandgap (EBG) structures [3]. However, the effects of EBG structures surrounding the antenna can be considered as two effect,

namely parasitic loading effect and cavity effect. The parasitic loading effect increases the bandwidth, whereas cavity effect is due to reflecting energy from EBG toward antenna results in a larger Q value and so decreasing the bandwidth. The EBG structure parameters and number of EBG rows is related to these effects.

In this paper, the influence of the EBG structures parameters and number of EBG rows on the far-field pattern of patch antennas is investigated. The changes in the far-field radiation patterns are discussed.

## 2. Theory of EBG

The parametric study on mushroom-like EBG structure is presented in [4]. It focused on four main parameters that affecting the overall performance of the antenna design. The parameters namely, patch width  $W$ , the spacing between mushroom-like EBGs, substrate thickness  $h$  and substrate permittivity  $\epsilon_r$ . In this paper, the study is focusing not only on  $W$ ,  $s$  and  $h$  as in [4], but also on the spacing between patch element,  $g$  and the number of rows of the EBG inserted between the patch elements.

Mushroom-like EBG consists of a ground plane, a dielectric substrate, metallic patches and vias that connecting the patches to the ground plane. The structure of this EBG and its equivalent lumped LC elements is shown in Figure 1. The inductance and capacitance of the circuit are due to the shorting vias and the spacing between the adjacent metal patches [5].

The central frequency of the band gap is

$$f = \frac{1}{2\pi\sqrt{LC}} \quad (1)$$

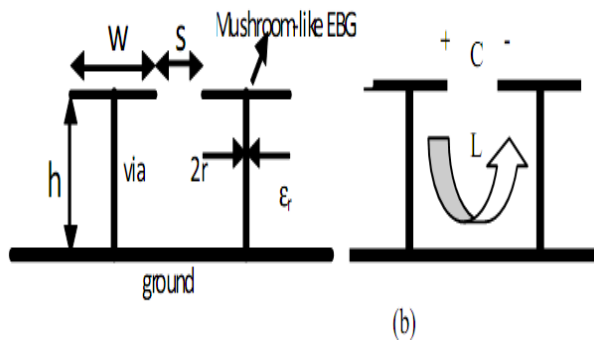


Fig.1 (a) Mushroom-like EBG structure (b) Lumped LC model.

### 3. Simulation of patch antenna integrated with EBG

The conventional microstrip antenna was designed on substrate (80\*80mm) having dielectric constant  $\epsilon_r = 2.5$  and height of the substrate  $h=1.588\text{mm}$ . The microstrip antenna is excited by a coaxial probe and the feed point is located at the distance ( $dx=1.7\text{mm}$ ) away from the edge of the patch. The length  $L$  and the width  $W$  have been taken as 8.3 and 11.34mm.

The antenna under investigation is a microstrip patch antenna integrated with one row of conventional mushroom like EBG patches located half wavelength ( $g=15\text{mm}$ ) far from antenna radiating edges in E-plane with resonant frequency at 10GHz figure 2. The parameters of EBG unit cell are:  $w$  (EBG patch width) =3.5mm,  $s$  (gap between adjacent patches) =1mm,  $r$  (radius of via holes) =0.2mm.

Figure 3 is shown return loss of antenna with and without EBG structure, and figure 4 shown E-plane pattern of these two antennas.

As shown in figure 3, as expected, bandwidth of antenna with one row EBG in E-plane is greater than antenna without EBG about 2%, due to domination of parasitic effects.

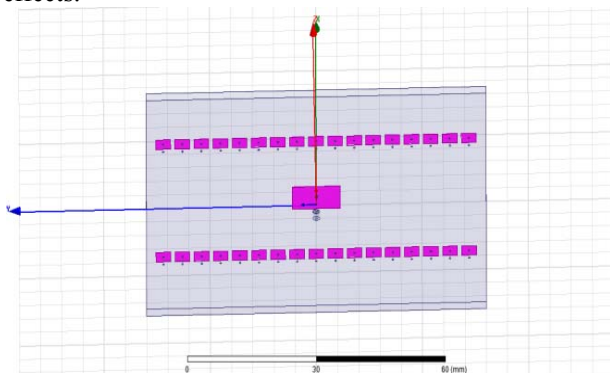


Fig. 2 Microstrip patch antenna with resonant frequency at 10 GHz

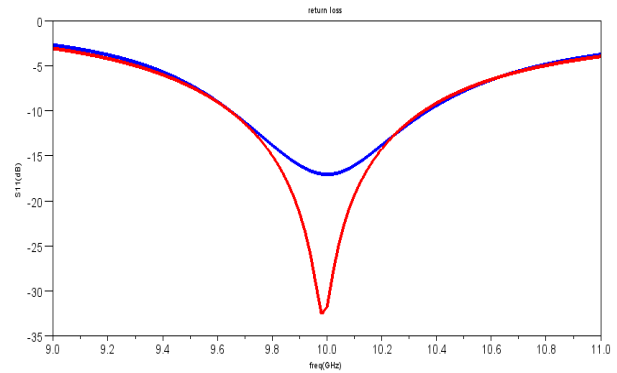


Fig. 3 Return loss of patch antenna with one row EBG (bleu) and without EBG structures (red)

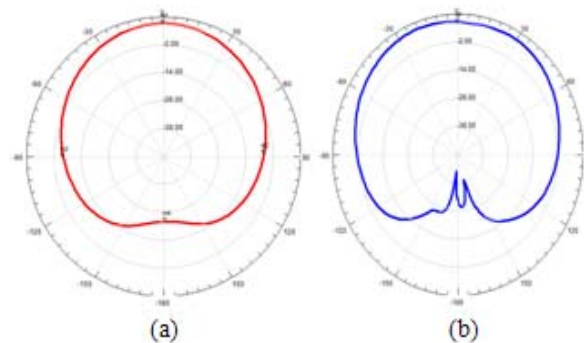


Fig 4 E-plane patterns of patch antenna (a) without EBG structures(b) with one row EBG

The reference antenna shows large radiation in the backward direction, and the antenna integrated with one row of conventional mushroom like EBG patches produces a lower backlobe, with less power wasted in the backward direction. Also, surface wave is reduced in EBG antenna.

In this part the antenna patch is simulated with increasing the number of EBG rows, figure 5 is shown the antenna patch with 4 rows of EBG structures in E-plane. Figure 6 is shown return loss of antenna with different number EBG rows.

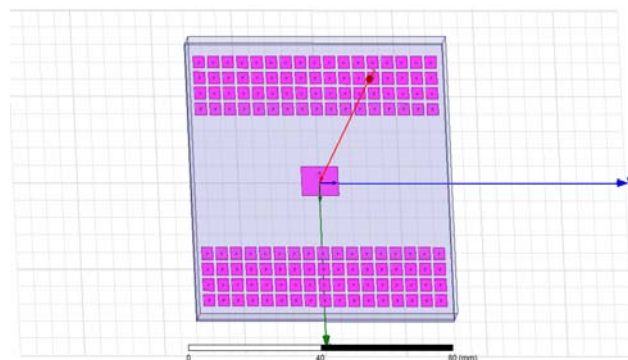


Fig. 5 Microstrip patch antenna, with 4 EBG rows.

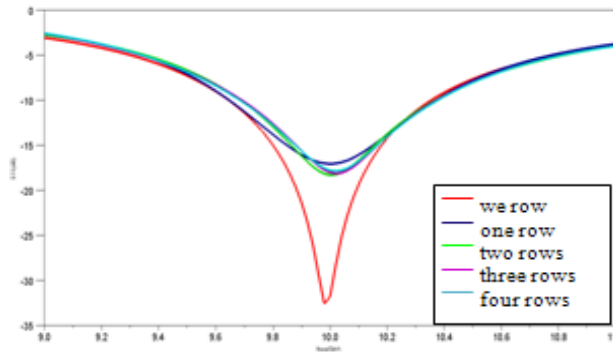


Fig.6 Return loss of the antenna with different number EBG rows.

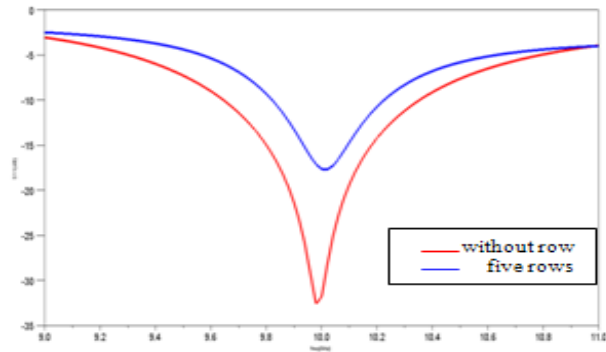


Fig.7 Return loss of the antenna with 5 EBG rows and without.

Table 1 shows the simulation results:

Table 1: Comparisons details between the results obtained with and without EBG rows

	<i>Res.fr eq</i> (GHz)	<i>S11</i> (dB)	<i>BP</i> (GHz)	<i>Gain</i> (dB)
We EBG	9.98	-32.61	0.708	7.27
One row	10	-17.06	0.721	6.64
Two rows	10	-18.30	0.68	6.80
Three rows	10.02	-18.13	0.68	6.20
Four rows	10.02	-17.85	0.68	6.05

The performance of the antenna without EBG row is about the same as the antenna with EBG rows, except that the return loss is dropped from -32.6 dB to about -18dB.

With increasing EBG rows from 2 to 4 rows, the bandwidth variation is negligible which is the indication of parasitic effects dominations.

The figure 7 is shown the return loss of antenna with 5 rows of EBG structures in the E-plane, figure 8 represented E-plane pattern for antenna with 5 rows. With 5 EBG rows, bandwidth suddenly decreases (0.708GHz to 0.401GHz) which is the indication of cavity effect domination.

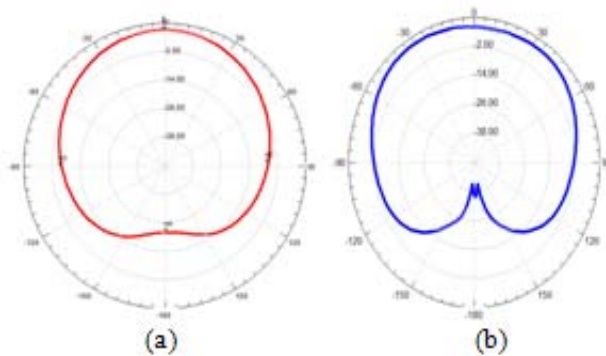


Fig.8 E-plane patterns of patch antenna (a) without (b) with 5 EBG row.

The reference antenna shows large radiation in the backward direction, and the antenna integrated with 5 rows of conventional mushroom like EBG patches produces a lower backlobe, with less power wasted in the backward direction. Also, surface wave is reduced in EBG antenna.

The same process is repeated in this part. The only different is that the spacing between patch is wider, that is 22.5mm (three quarter wavelength) from antenna radiating edges in E-plane to the row of conventional mushroom like EBG patches edge.

Figure 9 represented the return loss of antenna with one row.

It is seen from the Figure 9, the return loss for the conventional patch antenna is - 32dB at 10GHz and for the proposed patch antenna is - 21.25dB at 10.02GHz. A negative value for return loss shows that this antenna had not many losses while transmitting the signals.

With a wider spacing (three quarter wavelength), bandwidth suddenly decreases (0.708GHz to 0.64GHz) which is the indication of cavity effect domination.

The simulated results for gain that are obtained from conventional antenna and the proposed antenna on EBG substrates are shown in Figure 10.

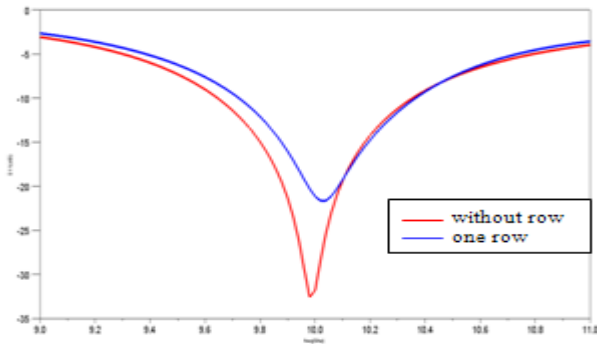


Fig.9 Return loss of the antenna patch with and without EBG row.

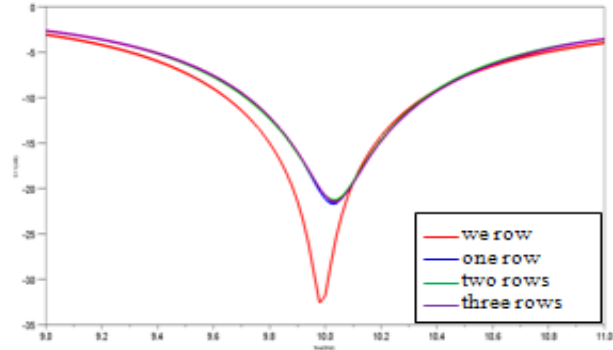


Fig.11 Return loss of the antenna with different number EBG rows.

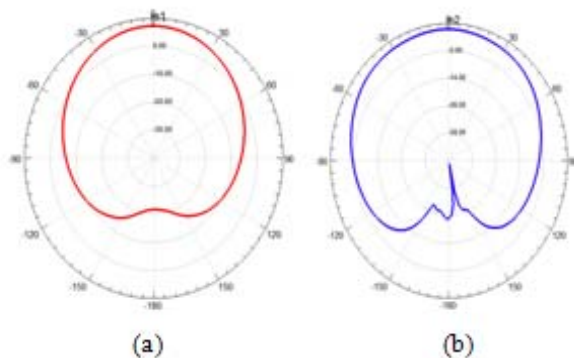


Fig.10 E-plane patterns of patch antenna (a) without EBG row (b) with one EBG row.

From the simulated results, it is shown that the gain of the conventional antenna and the proposed antenna is 7.26dB and 7.49dB. So, the gain of the proposed patch antenna on EBG substrates is 3% more than the conventional patch antenna.

Now, we increase the number of EBG row from 1 to 3, Figure 11 is shown return loss of patch antenna; table 2 shows the simulation results:

With increasing EBG rows from 1 to 3 rows, the bandwidth variation is negligible which is the indication of parasitic effects dominations, the gain of EBG patch antenna improved than the antenna without EBG. The conclusion from these two simulation results, the spacing from antenna radiating edges in E-plane to the row of conventional mushroom like EBG patches edge affect at the influence of side effects of EBG structure on the performance of the antenna.

Table 2: Comparisons details between the results obtained with and without EBG rows

	<i>Res.fr eq (GHz)</i>	<i>S11 (dB)</i>	<i>BP (GHz)</i>	<i>Gain (dB)</i>
We EBG	9.98	-32.61	0.708	7.27
One row	10.03	-21.64	0.642	7.48
Two rows	10.02	-21.27	0.637	7.42
Three rows	10.04	-21.47	0.640	7.4

#### 4. CONCLUSION

The patch antenna mostly used in modern mobile communication. The goals of this paper are to design conventional patch antenna and the patch antenna on EBG substrates with same physical dimensions that can operate at 10GHz and study the influence of the side effects of EBG structure on the performance of the antenna.

Based on the results obtained in this work, the following can be concluded from the parametric analysis; it is obvious that for 2 rows of EBG structures, an acceptable bandwidth is achieved. When 3 or 4 rows of EBG rows are used, bandwidth variations is negligible, but side and back lobe levels decreases in the cost of larger size consumption.

The spacing from antenna radiating edges in E-plane to the row of conventional mushroom like EBG patches edge controls the degree of influence of the side effects of EBG structure on the performance of antenna.

#### References

- [1] Jing Liang, and Hung-Yu David Yang “Radiation Characteristics of a Microstrip Patch over an Electromagnetic Bandgap Surface,” IEEE Transactions on Antennas and Propagation, Vol. 55, June 2007, pp.1691-1697.

- [2] Papalmerou, I., R.F. Drayton, and L.P.B. Katehi, "Micromachined patch antennas," IEEE Trans. Ant. Propag., Vol.46, No.2, pp.275-283, 1998.
- [3] Agi, K., K.J. Maloy, E. Schmiloglu, M. Mojahedi, and E. Niver, "Integration of a microstrip patch antenna with a two-dimensional photonic crystal substrate," Electromagnetic, Vol.19, pp.277-290, May Jun. 1999.
- [4] F. Yang and Y. Rahmat-Samii, "Electromagnetic Band Gap Structures in Antenna Engineering", 1<sup>st</sup> edition, Cambridge University Press, 2009.
- [5] M. Fallah and L. Shafai, "Enhanced Performance of a Microstrip Patch Antenna using a High Impedance EBG Structure", Antennas and Propagation Society International Symposium, Vol 3, 2003.

**BENIKHLEF Fethi** was born in Remchi, Algeria, in 1986. He received the Master degree in 2010, from the Abou Bekr Belkaid University in Tlemcen (Algeria). Since 2005, he joined the telecommunication Institute in Tlemcen University (Algeria). His research interests include contribution to the study of photonic band gap antennas.

**Noureddine Boukli-Hacene** was born in 1959 in Tlemcen, Algeria. In 1959, he received his Phd degree from the University in Toulouse (France), in electrical engineering from Limoges University, France, in 1985. Recently, he is a lecturer at the University of Tlemcen. His research interests include, among others, microstrip antennas and microwave circuits.

# Intelligent Video Object Classification Scheme using Offline Feature Extraction and Machine Learning based Approach

Chandra Mani Sharma<sup>1</sup>, Alok Kumar Singh Kushwaha<sup>2</sup>, Rakesh Roshan<sup>3</sup>, Rabins Porwal<sup>4</sup> and Ashish Khare<sup>5</sup>

<sup>1,3,4</sup>Department of Information Technology, Institute of Technology and Science  
Ghaziabad, U.P., India

<sup>2</sup> Department of Computer Engg. and Application, G.L.A. University,  
Mathura, U.P., India

<sup>5</sup> Department of Electronics and Communication, University of Allahabad,  
U.P., India

## Abstract

Classification of objects in video stream is important because of its application in many emerging areas such as visual surveillance, content based video retrieval and indexing etc. The task is far more challenging because the video data is of heavy and highly variable nature. The processing of video data is required to be in real-time. This paper presents a multiclass object classification technique using machine learning approach. Haar-like features are used for training the classifier. The feature calculation is performed using Integral Image representation and we train the classifier offline using a Stage-wise Additive Modeling using a Multiclass Exponential loss function (SAMME). The validity of the method has been verified from the implementation of a real-time human-car detector. Experimental results show that the proposed method can accurately classify objects, in video, into their respective classes. The proposed object classifier works well in outdoor environment in presence of moderate lighting conditions and variable scene background. The proposed technique is compared, with other object classification techniques, based on various performance parameters.

**Keywords:** Object Classification, Machine Learning, Real-time Video Processing, Visual Surveillance.

## 1. Introduction

Object classification is an important step in object detection [1], object activity recognition[2], content based video retrieval, visual surveillance[3], etc. Object classification techniques can be categorized primarily into two groups; Shape-based object classification [3-5] and Motion-based object classification [6-8]. The concept of Machine Learning can be used for classification and the training in such classification system is either based on pixels or features. The feature-based approaches are

preferred over pixel-based approaches, for object classification task, because encoding of ad-hoc domain knowledge using features is easier and it is difficult to train finite quantity of data using pixels [1]. Moreover, the feature-based object classification methods are better than the pixel-based methods in terms of speed. Haar-like features have been used, in synergy with different training techniques, for creating machine learning based classifiers [10-11]. Sialat et al.[10], in their pedestrian detection system, used Haar-like features along with the *decision tree*. A versatile object recognition technique has been proposed in [11], using multi-axial Haar-like features and a compact cascaded classifier. Viola et al.[1] also used the modified reminiscent of Haar-basis functions [12], for accomplishing object detection task.

The number of available features, for training the classifier, may be considerably high. Not all the features have equal importance, therefore some sort of mechanism is used to extract those features which are more important from classification point of view. Boosting combines different weak classifiers to form highly accurate predictors [13]. According to [14] AdaBoost is the best available binary classifier. Viola and Jones [1] proposed an object detection framework based on Haar-like features and AdaBoost, and used this framework for detecting the frontal human faces in real-time. One notable contribution, of the work of Viola et al.[1], was the concept of Integral Image, for fast extraction of Haar-like features. Since they have only two classes in their face detection case: Human and Non-Human. So they used AdaBoost for fast training of features. AdaBoost has a serious limitation that it can only be used to solve binary classification problem. AdaBoost requires that the accuracy of constituent classifiers be more than 50%. This condition, in AdaBoost,

is not easy to fulfill in case if number of the classes, in classifier, exceeds two.

This paper proposes a multiclass object classification technique based on Haar-like features and uses a Stage-wise Additive Modeling using a Multiclass Exponential loss function (SAMME) [13]. SAMME was originally used to classify the chemical data into multiple classes in [13]. This paper reports the classification results, obtained by the application of SAMME on image data. We use Integral Image representation, for fast feature evaluation and then multiple weak classifiers are trained using SAMME. The observation weights, associated with the feature points, are changed depending on their accuracy of classification during training process. Finally, the classifiers are linearly combined to form a strong classifier. The rest of the paper is organized as follows: Section 2 explains the multiclass boosting of classifier training, Section 3 describes the methodology of the proposed object classification technique, experimental results have been described in Section 4 and conclusions are given in Section 5.

## 2. Multiclass Boosting of Classifier Training

Objective of classification is to distribute the input vector (like feature points) into a set of  $K$  classes  $C_k$ , where  $k=1, \dots, K-1$ . These classes are disjoint so that each vector point  $x$  can be assigned to one and only one class. Boosting is an approach, used to improve the performance of classifier-training. Using boosting, the weak classifiers are trained sequentially and are finally merged, to form a classifier with adequate confidence of performance. A weight is assigned to every feature-point, to be classified, before classification. The feature points gain or lose weights during this process. Points, which are classified with error, gain weight and the points, which are correctly classified, lose weight. Points, with the higher weight value, seek more attention in the next level of classification, in order to classify them in the right classes. There are various classifiers [13- 15] available, which are created using this approach. AdaBoost [15] is a good approach for binary classification and there are two requirements associated with the classifiers in the AdaBoost- (i) the classifier must have an accuracy greater than 50% and (ii) the classifiers should be capable of representing the weighted data points. For the first condition, if the achieved accuracy is exactly 50% then distribution weights will not be updated. if the accuracy is less than 50% then the updation of distribution weights will take place in opposite direction. For binary classification, the random accuracy of classification of a feature point, in one of the two classes, should be at least 50% [15]. The second condition can be bypassed easily if the samples are taken from the training data set with replacement according to the weight distribution and then

passing to the component classifier. Suppose we try to use AdaBoost with a multiclass classifier, having three classes, then random accuracy will be 33.33%, which violates the first requirement for the use of AdaBoost. This constraint can be solved using two popular approaches [16]- (i) *one-against-all* and (ii) *one-against-one*. In case of *one-against-all* approach a separate model for each class is trained to distinguish the samples of that class from the samples of the remaining classes. For a data point to be classified using *one-against-all* approach, the class which gets the highest class prediction from the probabilistic binary classifiers, is assigned for that data point. In case of *one-against-one* approach, a classification model for each pair of classes is created. If the number of classes are  $K$  then there will be  $K(K-1)$  such models. Here, the class for a data point is decided by the voting classifiers in ensemble. The performance of such model based approaches is hampered by the management of models. Multiple models cause the processing speed to slow down. Therefore these approaches are not efficient options, to be used, for real-time visual surveillance applications. The problem of multiclass boosting can be solved by transforming it into several two-class problems. The general approach based on this concept is called Error-Correcting Output Coding (ECOC) [17]. ECOC is a method of making the most of the transformation of the multiclass problem into several binary problems. A simple ECOC method is known as Hamming Encoding. The problem of multiclass boosting of classifiers can be nicely solved with the SAMME [13], without creating any extra model and experimental results show that the execution speed is at par with binary AdaBoost based methods.

## 3. The Proposed Method

In proposed technique, the *Integral Image* representation is used for fast feature evaluation and boosting of cascaded classifiers is performed using SAMME. A more robust classifier is created by linearly combining the multiple weak classifiers. For experimentation, we consider three classes: *Human*, *Car* and *Non-Human-Car*. After classifier training, the objects, in the video, are classified into *Human* and *Car* classes. The proposed method has following steps:

### 3.1 Sample collection

The sample images for training the classifier are collected first. We have collected images for three classes- humans, cars and images which belong neither of these two from our own captured images and images from standard datasets like CalTek101, MIT-CMU datasets. We have created our own data set which consists of 4,000 images of humans, 3,500 images of cars and 5,000 images which are neither humans nor cars. These images were resized to



dimension 60x60 that they consist of only one object per image. This was performed in order to make the classifier learn more domain information from small number of images and this helps to improve the accuracy of the classifier.

### 3.2 Integral Image and Haar-basis function

We use simple Haar-like features which are reminiscent of Haar-basis functions as used in[1]. The use, of features instead of pixels, makes the classifier system work faster and helps in encoding the domain knowledge with finite quantity of data. We use three types of features- *two-rectangle feature*, *three-rectangle feature*, and *four-rectangle feature*. The Difference, between sums of pixels, gives the value of two-rectangle feature. The regions are horizontally or vertically adjacent and have the same size and shape (shown in Fig.1). A *three-rectangle feature* is used to compute the sum within two outside rectangles, subtracted from the sum in a center rectangle and a *four-rectangle feature* computes the difference between diagonal pairs of rectangles.

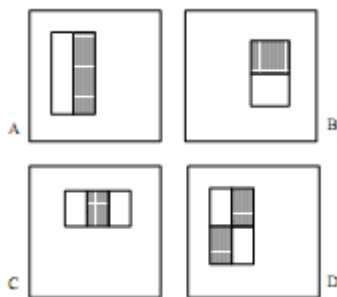


Fig1: Rectangle features

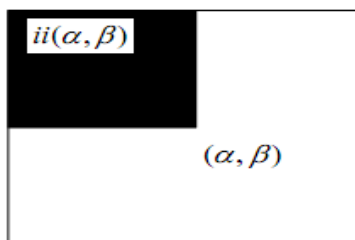


Fig2: Integral Image

The number of features, in a small detection window, can be considerably very high. For example, a base detection window of size 25x25 contains more than 200,000 such rectangle features. This huge number of features is very difficult to compute. To speed up this Computation process, we use *Integral Image* representation originally proposed by Viola et al. [1]. The Integral Image,  $ii(\alpha, \beta)$ , shown in Fig.2, at location  $(\alpha, \beta)$  contains the sum of all pixels above and left of  $(\alpha, \beta)$  and can be computed in a

single pass over image using the following pair of equations.

$$\omega(\alpha, \beta) = \omega(\alpha, \beta - 1) + i(\alpha, \beta) \quad (1)$$

$$ii(\alpha, \beta) = ii(\alpha - 1, \beta) + \omega(\alpha, \beta) \quad (2)$$

where,  $\omega(\alpha, \beta)$  is the cumulative row sum and  $i(\alpha, \beta)$  is the original image.

### 3.3 Multiclass boosting with SAMME

We use Stage-wise Additive Modeling using a Multiclass Exponential loss function (SAMME) [13] for boosting the cascade of multiclass classifiers. Viola et al.[1] used binary AdaBoost for their face detection system which fails in case of multiclass classification because of certain constraints over classifier's accuracy. The proposed method has three classes of classification and a modified AdaBoost is used for this. SAMME is an extension of AdaBoost with a modification. The key for applying SAMME for multiclass problem is that the component classifiers are no longer required to achieve an accuracy greater than 50%, but instead need only it to be better than random guessing. Suppose we are given a set of training data  $\{(\sigma_1, c_1), \dots, (\sigma_n, c_n)\}$ , where the input vector  $\sigma_i \in \mathbb{R}^p$ , and the  $c_i$  qualitatively assumes values in a finite set  $\{1, 2, \dots, K\}$  where  $K$  is the number of classes. The training data are independently and identically distributed samples from an unknown probability distribution. The goal is to find a classification rule  $\phi(\sigma)$  from the given training data, so that for a given new  $\sigma$ , we can assign it to a class label  $c$  from  $\{1, \dots, K\}$ . A complete SAMME multiclass boosting algorithm is given as below-

#### SAMME Multiclass Boosting Algorithm

- $\theta$  is vector of observation weights and  $\lambda$  is the vector length
  - $M$  is the number of stages in classifier.
  - $\alpha(\sigma)$  denotes a weak classifier
  - $\phi(\sigma)$  denotes the final classifier, created using the linear combination of many weak classifiers
  - $\eta^m$  is the score assigned to  $m^{\text{th}}$  classifier.
1. Initialize the observation weights  $\theta_i = \frac{1}{\lambda}$ ,  $i=1, 2, \dots, \lambda$ .
  2. For  $m=1$  to  $M$ 
    - (a) Fit a classifier  $\alpha^{(m)}(\sigma)$  to the training data using weights  $\theta_i$
    - (b) Compute the misclassification error rate as,
    - (c) Compute

$$\eta^m = \log \frac{1 - \text{err}^{(m)}}{\text{err}^m} + \log(K - 1) \quad (3)$$

$$(d) \text{ Set } \theta_i \leftarrow \theta_i \cdot \exp(\eta^m \cdot \delta_{(C_i \neq \alpha^m(\sigma_i))}) \quad (4)$$

for  $i=1, 2, \dots, \lambda$ .

(e) Re-normalize the value of  $\theta_i$

### 3. Output

$$\phi(\sigma) = \arg \max_k \sum_{m=1}^M \eta^m \cdot \delta(\alpha^m(\sigma) = k) \quad (5)$$

Here, weak classifiers are linearly combined to form a strong classifier.

### 4. End

The above boosting algorithm is similar to AdaBoost but with a small but crucial difference in Eq. 3. The term  $\log(K-1)$  has been added in the equation. From this term, it is evident that at  $K=2$ , SAMME reduces to AdaBoost but in multiclass case ( $K>2$ ) the term  $\log(K-1)$  is critical. One of the benefits is that for  $\eta^m$  to be positive, we just need have  $(1-\text{err}(m))>1/K$  or the accuracy of each classifier to be better than the random accuracy rather than 50%. In the present case, we have three classes- *Human*, *Car* and *Non-Human-Car*. Here, the accuracy of the weak classifiers needs to be only more than 33.33%, rather than being more than 50%. SAMME behaves as the forward stage-wise additive modeling for multiclass classification case.

## 4. Experimental results

The proposed object classification technique has been tested on several videos, captured in real outdoor environment. The results, with one such representative video, are given in Fig.3 using the proposed approach. In these results, red windows are used for classifying the human objects and blue windows are used to classify the cars. Fig.3 shows the object classification results with natural lighting conditions.

One can observe from the results presented in Fig. 3 that there are three human objects and one car in the video. The video was shot at a frame resolution of  $640 \times 480$  and frame rate of 20 frames per second. The video consists of total 1600 frames and we have posted results at difference of 25 frames. In the starting of video the humans start moving towards camera and later on occupy the random motion paths. The car shown in the video has been parked and does not move in video. The proposed technique is capable of handling the partial inter-class and intra-class occlusions. Video frames no. 50, 225, 250 and 275 have

the partial human to human occlusion and frames no. 1025, 1050 and 1400 show the occlusion between car and human. The proposed method classifies the objects even in these frames accurately and in case of full occlusion proposed method has the capability of resuming quickly. As it is clear from the results given in Fig.3 that the human objects have the varying poses of their bodies as they move in the video in appearing in all the possible natural poses. Moreover, they acquire various views like frontal, side and back to the camera. Also, their speed of movement varies. Unlike various existing classification methods, the proposed method classifies the various objects accurately irrespective of their views of appearance.

Many existing object classification schemes require the lighting conditions to be good in the video and can work only in presence of the statically good light. But from practical point of view it is not desirable. As in case of visual surveillance the videos are captured in the outdoor environment and fluctuations in the lighting are unavoidable. The classification results using proposed technique in poor lighting condition are given in Fig. 4. The video for this experiment was shot in evening at University of Allahabad campus. The resolution of video frames is  $640 \times 480$  and was shot at frame rate of 20 frames per second. The video consists of a total 800 frames. The classification results with this video are given below starting from frame number 25 to frame number 800 at a difference of 25 frames. It can be observed that in the video three humans and one car appear. Humans are walking in the random directions. Car has been parked and does not move. The humans while walking cause occlusion within themselves and with car. Frames 250 and 675 show the partial occlusion between humans where one human is partially occluded from another. Frames 625, 650, 725, 750, 775 and 800 show the occlusion between car and human. In frame number 625 and 650, one human object occludes the car while in frame number 725, 750, 775 and 800 two humans appear before car and occlude it. It is evident from the results that the proposed technique accurately classifies the objects in all the aforementioned frames also.

We have tested the proposed method on several other realistic videos shot in outdoor environment. The average execution speed on these videos has been observed as 23 frames per second. The detection accuracy of the proposed classifier system has been estimated in our experiments between 91% and 98%. The tradeoff between the accuracy and the execution speed can be customized easily depending upon the user's requirement. The execution speed can be increased by slightly decreasing the detection accuracy and the detection accuracy can be increased by a slight decrement in execution speed.



frame 25



frame 50



frame 275



frame 300



frame 75



frame 100



frame 325



frame 350



frame 125



frame 150



frame 375



frame 400



frame 175



frame 200



frame 425



frame 450



frame 225



frame 250



frame 475



frame 500



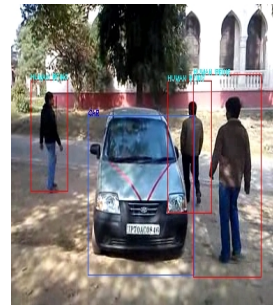
frame 475



frame 500



frame 725



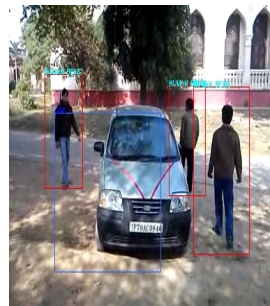
frame 750



frame 525



frame 550



frame 775



frame 800



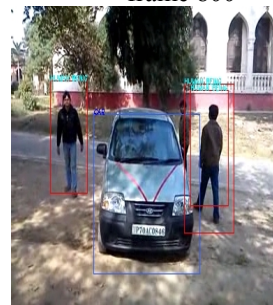
frame 575



frame 600



frame 825



frame 850



frame 625



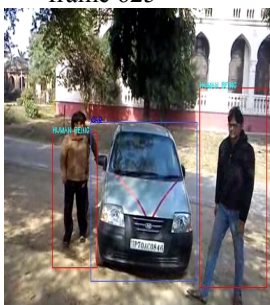
frame 650



frame 875



frame 900



frame 675



frame 700



frame 925



frame 950



**Fig. 3:** Object classification with the proposed technique in normal lighting condition.

The operation of the proposed method has also been tested with the implementation of a real world system. The system was kept under test operation for 60 hours continuously. The system was configured with a single mounted camera. The system processed the captured video stream in real-time without any time delay. Object classification produced adequately accurate results.

#### 4.1 Comparison of object classification methods

We compare the proposed method of object classification with various other existing state-of-the-art methods. This comparison is shown in table 1. Chosen parameters for comparison are: number of classes, number of training samples, object classification accuracy, processing speed, method used for feature extraction and machine learning algorithm used for classifier-training. This performance statistics has been observed on the uniform hardware set up. The performance of the proposed technique is comparable to other existing techniques. The peak detection accuracy of our method is 98% which is more than the peak accuracy of the object classification methods [19-22]. It can be observed from the table 1 that the computational speed is also comparable to the other methods. The proposed method has computational speed of 25 frames per second which is good enough for real-time processing of video frames.

Object classification is a very crucial step towards object activity recognition and behavior prediction. If done accurately then it can considerably improve the quality and accuracy of the various applications such as in monitoring and visual surveillance systems, content based video retrieval and indexing, access control to restricted areas and crowd flux analysis. Machine learning based approaches for object classification are gaining much popularity compare to the other approaches of object classification methods. This is because of that a huge amount of unannotated training data is readily available for the unsupervised training of these systems.

Also, various standard annotated training sets are available including a range of object classes in order to facilitate the supervised training of the system as well. The customizable applications for this purpose have started appearing which include training data set as well a user friendly kit to train with that data set. Generally, these types of kits are able to detect only a very few number of object classes. Majority of them can only recognize (distinguish) only one class of object. Generally these systems are known as object detectors (contrast with object classifiers) and are able to detect, generally, only a single type of objects out of various types of objects such as human detector, car detector, cow detector etc. These detector systems may also be trained to detect some special part of an object instead of recognizing the object as a whole e.g. Viola et al.[1] developed a human face detector system which was capable of tagging the human faces in an image. The system is slower and takes much time to detect the faces. Sun et al.[20] also detected the vehicles using their vehicle detector system. This system is even slower than the system of Viola et al.

**Table1:** Comparison of the proposed method with other existing methods for object classification

Method Name	Criteria for Evaluation				
	Number of Classes and Training Samples	Accuracy	Speed	Method used for Feature Extraction	Machine Learning Algorithm used
Kato et al. [19]	2 Classes: 5000 Vehicle and 5000 Non-Vehicle samples to train	89.0% to 96.0%	0.5- 1.7 Sec per Frame	Not Specified	MC-MQDF - Linear Classifier
Sun et al. [20]	2 Classes: 2500 Vehicle, 2500 Non-Vehicle Samples to train	88.9% to 96.4%	1.0-2.0 Sec per Frame	PCA, Wavelet, Gabor Features	Neural Networks & Support Vector Machines
Viola et al. [1]	2 classes: 2000 Face and 3000. Non-Face \Samples to train	78.3% to 98%	0.07- 0.5 Sec per Frame	Haar-like features	AdaBoost
Opelt et al. [21]	3 Classes: 450 Person, 350 Bike 250 Non Bike / Person Samples to train	65.0 to 83.5%	2.1-3.0 Sec per Frame	Intensity Moments, SIFTs	AdaBoost
<b>Proposed Method</b>	3 classes: 4000 Human, 3500 Car and 5000 Non-Car/Human Samples to train	91.6% - 98.3%	0.01 - 0.05Sec per Frame	Haar-like Features	SAMME

#### 5. Conclusions

A real-time multi object classification approach has been proposed in this paper. The approach first trains a multiclass classifier using Haar-like features and SAMME boosting strategy. The concept of *Integral Image* is used for fast feature evaluation. The proposed technique is a

good substitute for AdaBoost and can work beyond the limitations of AdaBoost in multiclass environment.

The proposed method has a peak detection accuracy of 98.30% and it can process 25 frames per second. The comparison, with various other existing object classification methods, proves the novelty of the method. Experimental results show that the proposed technique works fine even in poor lighting conditions and in the presence of partial occlusion.

## References

- [1] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in Proc. IEEE Int. Conf. on Computer vision and Pattern Recognition, vol. 1, pp. 83-87, 2001.
- [2] C. M. Sharma, A.K.S. Kushwaha, S. Nigam and A. Khare, "Automatic human activity recognition in video using background modeling and spatio-temporal template matching based technique," in Proc. ACM Int. Conf. on Advances in Computing and Artificial Intelligence, pp.97-101, India, 2011.
- [3] W. Hu and T. Tan, "A survey on visual surveillance of object motion and behaviors," IEEE Trans. on Systems, Man, and Cybernetics, vol. 34, no. 3, pp. 334-352, 2006.
- [4] R. T. Collins, A. J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, P. Burt, and L. Wixson, "A system for video surveillance and monitoring," Carnegie Mellon Univ., Pittsburgh, PA, Tech. Rep., CMU-RI-TR-00-12, 2000.  
[http://www.ri.cmu.edu/publication\\_view.html?pub\\_id=3325](http://www.ri.cmu.edu/publication_view.html?pub_id=3325)
- [5] A. J. Lipton, H. Fujiyoshi, and R. S. Patil, "Moving target classification and tracking from real-time video," in Proc. IEEE Workshop on Applications of Computer Vision, pp. 8-14, 1998.
- [6] Y. Kuno, T. Watanabe, Y. Shimosakoda, and S. Nakagawa, "Automated detection of human for visual surveillance system," in Proc. Int. Conf. on Pattern Recognition, pp. 865-869, 1996.
- [7] R. Cutler and L. S. Davis, "Robust real-time periodic motion detection, analysis, and applications," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 22, pp. 781-796, 2000.
- [8] A. J. Lipton, "Local application of optic flow to analyze rigid versus non-rigid motion," in Proc. ICCV Workshop on Frame-Rate Vision, Corfu, Greece, September 1999.  
<http://www.amazon.co.uk/application-non-rigid-Technical-University-Institute/dp/B0006RFKWW>
- [9] C. Stauffer, "Automatic hierarchical classification using time-base co-occurrences," in Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, vol. 2, pp. 335-339, 1999.
- [10] M. Sialat, N. Khlifat, F. Bremond, and K. Hamrouni, "People detection in complex scene using a cascade of boosted

classifiers based on Haar-like Features," in Proc. IEEE Int. Symposium on Intelligent Vehicles, pp. 83-87, 2009.

- [11] J. Nishimura and T. Kuroda, "Multi-axial haar-like features and compact cascaded classifier for versatile recognition," IEEE Sensors Journal, vol. 10, no. 11, pp. 1786-1795, 2010.
- [12] C. Papageorgiou, M. Oren, and T. Poggio, "A general framework for object detection," in Proc. IEEE Int. Conf. on Computer Vision, pp. 555-562, 1998.
- [13] J. Zhu, H. Zou, S. Rosset, and T. Hastie, "Multiclass adaboost," Int. J. of Statistics and Its Interface, vol. 2, pp. 349-360, 2009.
- [14] L. Breiman, "Bagging predictors," Int. J. of Machine Learning, vol. 24, pp. 123-140, 1996.  
<http://portal.acm.org/citation.cfm?id=231989>
- [15] Y. Freund, and R. Schapire, "A decision theoretic generalization of on-line learning and an application to boosting," J. of Computer and System Sciences, vol. 55, 119-139, 1997.
- [16] M. Rodriguez, "Multi-class boosting," 2009.  
[http://amachinelearningtutorial.googlecode.com/files/boosting\\_per.pdf](http://amachinelearningtutorial.googlecode.com/files/boosting_per.pdf)
- [17] I. H. Witten, and E. Frank, "Data mining: Practical machine learning tools and techniques," 2005.  
<http://200.201.9.37/mestrado/docs/WittenFrank.pdf>
- [18] A. Laika and W. Stechele, "A review of different object recognition methods for the application in driver assistance systems," in Proc. IEEE 8th Int. conf. on Image Analysis for Multimedia Interactive Services, 2007.  
<http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4279117>
- [19] T. Kato, Y. Ninomiya, and I. Masaki, "Preceding vehicle recognition based on learning from sample images," IEEE Trans. on Intelligent Transportation Systems, vol. 3, no. 4, pp. 252-260, 2002.
- [20] Z. Sun, G. Bebis, and R. Miller, "Monocular pre-crash vehicle detection: features and classifiers," IEEE Trans. on Image Processing, vol. 15, no. 7, pp. 2019-2034, 2006.
- [21] U. Handmann, T. Kalinke, C. Tzomakas, M. Werner, and W. V. Seelen, "An image processing system for driver assistance," Int. J. of Image and Vision Computing, vol. 18, no. 5, pp. 367-376, April 2000.
- [21] A. Opelt, A. Pinz, M. Fussenegger, and P. Auer, "Generic object recognition with boosting," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 28, no. 3, pp. 416-431, 2006.

**Chandra Mani Sharma** received his B.Sc. from Meerut College in 2006, MCA in year 2009 from ITS Ghaziabad and M.Tech. in Computer Science from Devi Ahilya University, Indore in year 2011. He worked as a research fellow on UGC-Sponsored research project *An Intelligent Video Surveillance System for Human Behavior Analysis* at IPCV Labs, University of Allahabad from May, 2010 to June, 2011. He has obtained various awards and scholarships including ITS-meritorious scholarship, IETE scholarship and MHRD-India scholarship. Currently, he is holding the position of assistant professor in IT at ITS Ghaziabad, India.

**Alok Kumar Singh Kushwaha** received his B.Sc. and M.Sc. from University of Allahabd. He has done his M.Tech in Software Engineering from Devi Ahilya University, Indore in 2011. He is working as an assistant professor in Computer Science at GLA University Mathura.

**Rakesh Roshan** received his B.Sc. degree from Ranchi University, MCA from UPTU Lucknow, M.Tech (IT) from Mysore and he is pursuing his Ph.D from MBU, Solan. His area of interest includes artificial intelligence, data bases, and software engineering. He is IBM DB2 and RAD certified professional. Currently he is working as an assistant professor in IT at ITS Ghaziabad, India.

**Dr. Rabins Porwal** earned his B.Sc. degree from CSJMU Kanpur, M.Sc. and Ph.D from DEI Agra. He is senior fellow of IEEE. He got various awards and scholarships during his academic and research career including JRF, SRF, CSIR fellowship, gold medal in graduation, national scholarship from 10<sup>th</sup> standard up to graduation etc. He is treasurer of Computer Society of India (Ghaziabad section). He is life member of CSI and The Indian Science Congress Association (ISCA). He is a member of ACM and Indian Society for Technical Education (ISTE). He is currently working as an associate professor in IT at ITS Ghaziabad, India.

**Dr. Ashish Khare** received his B.Sc, M.Sc and Ph.D degrees from University of Allahabad. Dr. Khare has more than 50 research papers in international/national journals and conference proceedings. His research area includes image processing, medical imaging, video processing, soft computing, and algorithm analysis. He is currently employed as an assistant professor at Department of Electronics and Communication, University of Allahabad.



# An Extensible and Secure Framework for Distributed Applications

Aneesha Sharma<sup>1</sup>, Shilpi Gupta<sup>2</sup>

<sup>1</sup>Department of Computer Science and Engineering, Amity University  
Noida (201301), Uttar Pradesh, India

<sup>2</sup>Department of Computer Science and Engineering, Amity University  
Noida (201301), Uttar Pradesh, India

## Abstract

Availability, Scalability, Reliability, Security and resource sharing are the key issues for success of any application, that are well addressed by distributed applications. Distributed applications provide services to different computers located at various locations that are connected by some means of communication network. In distributed systems a particular site consists of various computing facilities and an interface to local users and to a communication network. This paper provides various issues that must be taken into consideration while developing distributed systems. The issues discussed in this paper offer a secure framework for developing any distributed application on the top. Of these issues there are certain most commonly occurring issues that a distributed system fall victim to.

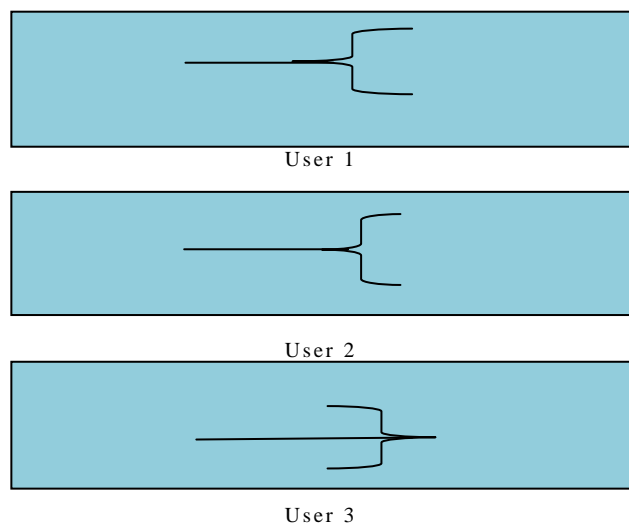
**Keywords:** distributed system, message passing, consensus, clock Synchronization, deadlock detection, concurrency management.

## 1. Introduction

A Distributed System is viewed as a set of computers that are independent in nature and in which each computer has its own local memory, operating system and clock[1]. The means of communication in a distributed system is through message passing method. In message passing method there is a sender and a receiver. A sender communicates with the receiver by means of passing messages. The information to be shared is copied from the sender process's address space to the address space of all the receiver processes, and this is done by transmitting the data to be copied in the form of messages [2]. There are several issues that must be considered while developing a distributed system these include message passing, deadlocks, concurrency etc. All of these and many more will be discussed in detail in this paper.

## 2. Problem Statement

As we know that distributed system is widely used in building many applications. So, through this paper an idea is given to build a Multi User Virtual Drawing Board (MUVDB) that uses the concept of distributed systems. Multi user virtual drawing board can support multiple users at the same time simultaneously and all the changes made by each user are reflected in all the virtual boards. The aim is to increase availability and provide scalability. Multi user virtual drawing board for extensible distributed framework is basically a drawing application that can support many users simultaneously so that all can work to achieve a common task considering views and ideas of all the users dedicated to perform the same task to design an application that is best in its design.



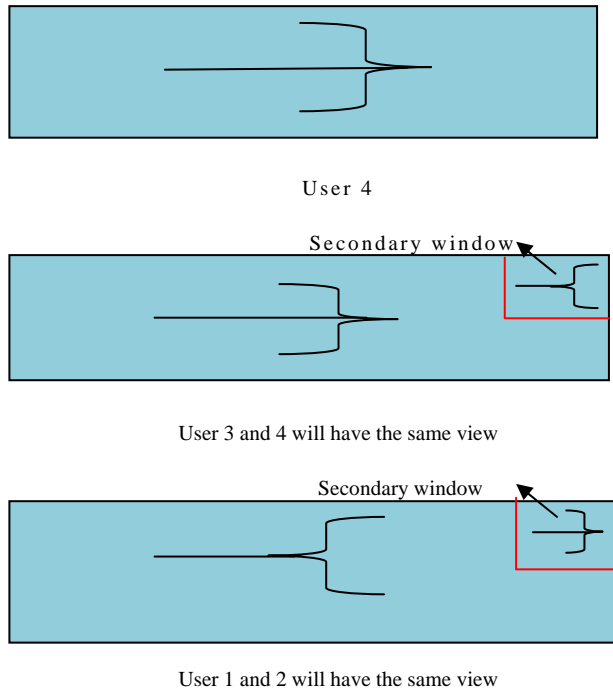


Fig. 1 interface design for MUVDB

The interface for the Multi User Virtual Drawing Board will look in the same manner as shown through the fig. 1. The users working on a design are free to give their respective views on a particular part of the design. The users that agree on the same design will be grouped together and a secondary window is also provided on the main window. This secondary window is provided in order to have the view of the designs made by other users. The number of secondary windows and groups can be controlled. The threshold can be decided by the number of users interested and supporting a particular design. The design instances can be discarded if no users are interested. The proposed idea of the MUVDB will be carried out in three parts. First a message passing model will be designed using sockets and threads. Secondly, an interface will be designed as shown in figure. 1 using swings. Lastly, concurrency or serialization protocol will be applied. Through this paper only an idea for building such an application is given and no implementation has been provided for the same. The paper is divided into various sections starting from introduction, problem statement, basic architecture, issues in distributed systems, conclusion and future scope. Through the issues in distributed systems, an attempt is made to compare the proposed idea with the various issues discussed in this paper. The aim is to state how these various issues can be handled in the multi user virtual drawing board and the basic advantage behind such an idea.

### 3. Basic Architecture

A distributed system consists of many computers located at different locations and all are connected via communication network. Each of these computers have their own local memory, operating system and clock apart from a Global clock. All of these clocks are synchronized in order to have effective error free communication. A machine at a particular site consists of two types of processes running on it. These processes are:

- i. Local process (LP)
- ii. Coordinator process (CP)

Communication between two machines in a network is carried out via Sockets. And communication between local process and a coordinator process is carried out via Inter Process Communication (IPC). Local process is basically used to draw an application and coordinator process helps the local process to see what is happening on other machines.

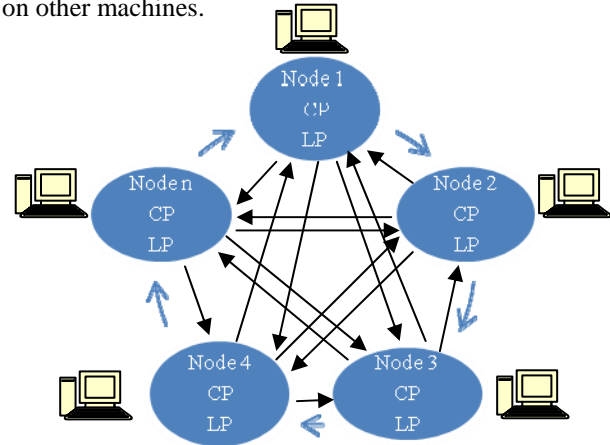


Fig. 2 Distributed Framework

### 4. Issues in Distributed Systems

Various issues in distributed systems are:

**4.1 Message Passing Framework-** In distributed systems there is a concept of inter process communication. There are two most important types of IPC's that are most commonly used these are shared memory model and message passing model. In both of these types of models failure may occur.

Conventional message passing technologies are:

**4.1.1 Unreliable datagrams-** identifies the corrupted messages among the stream of messages and then discards such messages. This technology generally fails because of its limitations to provide additional processing because of which most of the messages get through, some may get lost in transmission, duplicated or are delivered out of order [3].

4.1.2 *Remote procedure call (RPC)* - is said to be a reliable service. It works by providing communication by invoking a procedure that returns a result. But in this situation also failure may occur at the sender or at the receivers site. Failure may also occur in the network which may cause delay in the delivery of request or reply [3].

4.1.3. *Reliable data streams*- in this form of message passing technology communication is carried out over channels and these channels provide sequenced message delivery in a flow control and reliable manner [3]. ISIS is a system that provides various tools to support the construction of a reliable distributed system [3]. According to ISIS, Group communication involves various types of groups these are:

i. *Peer groups*- this type of scenario can be seen when the set of processes cooperate with each other, for example ,to replicate data [3].

ii. *Client-server groups*- in this type of group communication a process can communicate with any group provided the group name and permissions are given [3].

iii. *Diffusion groups*- are formed by a client-server group. In diffusion group clients register themselves and the members of the group send messages to the full client set and the clients are the passive sinks.

iv. *Hierarchical groups*- are built from multiple component groups, for the reason of scalability.

H. Attiya et al. [4] proposed a message passing model, in which various computations performed in the system are viewed as sequences of steps. In this model each step is of two types either a message delivery step delivering a message to the processor or a computation step of a single processor.

Message passing approach can be used in Multi User Virtual Drawing Board as the medium to exchange messages between number of users working on a common design and this can be achieved by building a message passing model using sockets and threads.

**4.2 Clock Synchronization-** distributed system is a collection of independent computers. The main reason behind the development of such a system is to achieve load balancing and resource sharing in the network. For this purpose it is necessary that the clocks of the communicating nodes should agree to a common clock value [1]. Now if the system is being employed to work for a real time application then it is must that all the clocks of the processors must match with Coordinated Universal Time, UTC.

Factors that cause errors in the clock are:

i. *Clock skew*- occurs when two clocks run at an exact same speed but have a constant difference.

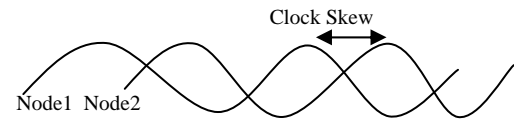


Fig. 3 Clock Skew

ii. *Drift rate*- occurs when the clocks do not run at an exact same speed. And this difference increases to a considerable level after some time and continues to be so.

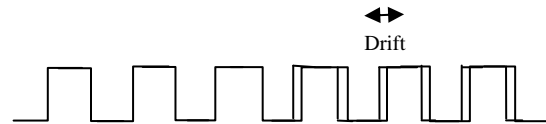


Fig. 4 Clock Drift

In the paper written by Kasim Sinan Yildirim [11], a tool is provided for finding the lower bounds of the distributed clock synchronization algorithms. And with the help of this tool the lower bound on the clock synchronization error between two processors in a distributed system can be proved.

In the method for clock synchronization proposed by Latha CA et al. the nodes in a distributed system are connected in the form of a ring. A Sync Token is a specific bit pattern that is made to rotate in the ring. Out of all the nodes in a ring one node is allowed to have a direct connection with the UTC server and this node is referred to as chief time server (CTS). In the beginning, CTS has the sync token and acts as a time server. It can then get the UTC value from UTC server. CTS then sets its clock value as the received UTC value and then broadcast that value to all the nodes connected in a ring [1].

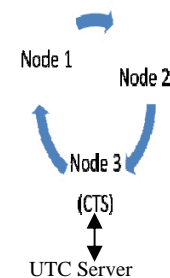


Fig. 5 Clock synchronization

4.2.1 *Event ordering*- for ordering of various events Lamport [7] defined a relation known as “happened before” and introduced the concept of logical clocks.

According to *happened before* relation on a set of events:

- If  $a$  and  $b$  are events in the same process, and  $a$  comes before  $b$ , then  $a \rightarrow b$ .
- If  $a$  is the sending of a message by one process and  $b$  is the receipt of the same message by another process, then  $a \rightarrow b$ .

- If  $a \rightarrow b$  and  $b \rightarrow c$ , then  $a \rightarrow c$ .

4.2.2 *Logical clock*- lamport [7] gave the concept of logical clocks. A clock say  $C_i$  is associated with each process  $P_i$  and this process assigns a number  $C_i(a)$  to any event  $a$  in that process. Logical clocks are a way to associate a timestamp to the events in a system. This allows the events to be properly ordered that are related to each other by the happened-before relation [2].

4.2.3 *Vector clock*- a vector clock system basically involves timestamping mechanism. Vector clocks were designed with the aim to allow processes to track the concurrency between the events produced by these processes [12]. The events that are produced by processes are mainly message sendings, message receives or internal events. A vector clock is defined as an array of  $n$  integers in which one entry is for one process for example if we take entry  $j$  it will tell us the number of events that are produced by the process  $p_j$ . Now the timestamp associated with each event defines the current value of the vector clock for a particular process that produced that event. With the help of these vector clocks it becomes easy to identify that whether two events are casually related or not.

Vector clocks are mostly preferred to achieve an extensible distributed framework like the Multi User Virtual Drawing Board because they provide global ordering and can solve various problems of casual broadcast, detection of message stability and detection of an event pattern [12]. A casual broadcast was first introduced by Birman and Joseph [12, 13].

The main problem of vector clock is scalability, R.Baldoni [12] in his paper presented a method to overcome this problem by giving the concept of Bounded Vector Clocks.

In the paper written by Li-Hsing Yen [14], a method for preserving the functionality of vector clock was given by performing correct clock resetting mechanism.

**4.3 Deadlock**- a deadlock is a situation where in the resources held by other processes are being requested by some other process in the same set [5].

Deadlock handling techniques:

4.3.1 *Deadlock prevention*- can be achieved by two ways:

→By allowing a process acquire all the resources it needs before it can actually begin execution. But this method involves various drawbacks like it effects the systems concurrency, may lead to certain processes enter into a deadlocked condition and also in certain systems it is difficult to actually predict what resources a process might need in future.

→By pre-empting the process that holds all the resources needed by a particular process. This method also have drawback like several processes are pre-empted without any deadlock.

4.3.2. *Deadlock Avoidance*- in case of distributed systems deadlock avoidance is achieved by granting resource to a process if the resulting system state is safe [5]. This includes various drawbacks and makes it impractical to be used in distributed systems:

→In order to keep account for the safe state of the system by every site, large storage capacity and wide communication capability is required.

→A condition might occur where in several sites perform the safe state checking but the resulting net state may not be safe.

→This method for performing check for safe state is computationally expensive when large number of processes and resources are involved.

4.3.3. *Deadlock Detection*-Involves studying the relationship between the process and the resource to identify the presence of cyclic wait [5]. In distributed systems deadlock detection is advantageous in comparison to deadlock prevention and deadlock avoidance due to following reasons:

→Detecting a cycle in the system does not hinders the usual activities of a system.

→When a cycle is formed in the state graph it remains there in the system. Until it is detected and broken.

Issues in deadlock detection:

- State graph maintenance.
- To search a state graph for the presence of cycles.

Deadlock detection algorithms are divided into three parts:

4.3.3.1. *Centralized deadlock detection algorithm*- in centralized algorithms for deadlock detection, a control site is there that maintains the state graph for the entire system and performs the checks for the existence of any deadlock cycle. All sites are capable to request or release resources by sending "request resource message" and "release resource message". The control site updates its state graph whenever it receives such a message. This algorithm is though simple and feasible to implement but it is inefficient because it requires all the messages to go all way to the control site which in turn causes long delays for user request, traffic problem near the control site. All these reasons make this algorithm unreliable because everything depends on the control site [5].

Ho-Ramamoorthy presented two centralized deadlock detection algorithm :

i. *Two-phase algorithm*- a status table is maintained by every site in order to record the status of all the processes initiated at that particular site. A state graph is constructed by a designated site by requesting the state table from all the sites and it is then checked for the presence of cycle. If no cycle is formed then the system is not deadlocked. Then this designated site again constructs a state graph using only the transactions common to both reports by requesting status table from

all sites. A system is declared deadlocked if the same cycle is detected again [5]. In this algorithm there may be a possibility of detecting a false deadlock so it is not that efficient.

ii. *One-phase algorithm*- involves only one phase of status reports from each site. However, two status tables are maintained by each site one for resource status and another for process status. The resource status table records all the transactions that have locked or are waiting for the resources stored at a particular site. The process status table keep a record of all the resources

locked by or the resources that are being waited by all the transactions at a particular site [5]. A state graph is constructed by a designated site, by requesting both the tables from all the other sites and by using only the transactions that are same for both the resource table and process table and then performs check for the cycle. This algorithm does not detect the false deadlocks as it eliminates any inconsistency in state information by using only the information common to both tables. The one-phase algorithm is efficient in comparison to two-phase algorithm.

Table 1: Deadlock detection algorithms in distributed systems

Basis for distinction	Centralized Algorithm	Distributed Algorithm	Hierarchical Algorithm
1.State graph	Maintained at a single site called control site.	Distributed over several sites.	Sites are arranged in form of hierarchy.
2.Implementation	Easy	Difficult	Intermediate
3.Deadlock resolution	Simple	Cumbersome	Deadlocks are localized to as few clusters as possible.
4.Single point of failure	Yes	No	Employs to get best of both centralized and distributed.
5.Example	Ho-Ramamoorthy algorithm	Goldman's algorithm, Isloor-Marsland algorithm etc.	Menasce-Muntz algorithm.

4.3.3.2. *Distributed deadlock detection algorithm*- the work of detecting the deadlock cycle is distributed to all sites as all the sites co-operate each other to detect a deadlock cycle.

Various distributed deadlock detection algorithms are:

i. *Goldman's algorithm*- involves the use of an ordered blocked process list (OBPL), for exchanging deadlock related information. In OBPL each process is blocked by its successor and the last process in it will be either waiting to get an access to a particular resource or will be in a running state [5]. In this algorithm deadlock is detected by expanding the OBPL by attaching the process that holds the resource that is needed by the last process in the OBPL list at the end and this process is continued until a deadlock is detected or the OBPL is removed. Advantage of Goldman's algorithm is that whenever deadlock detection is to be carried out then only OBPL list is constructed.

ii. *Isloor-Marsland algorithm*- is also known as "online" deadlock detection algorithm. This algorithm detects deadlocks at the time of making decisions about resource allocation at a particular site. A reachable set for a particular node is constructed that consists of all the nodes that can be reached from it. A specific process is deadlocked if in a reachable set of a particular node, consists of that node also in the set. Deadlocks are detected by constructing reachable sets for all nodes. Each and every site maintains a state graph for the

system and also the reachable sets for each node in the state graph.

iii. *Obermarck's algorithm*- provides a method to detect multisite deadlocks without the necessity to maintain a huge global transaction wait for graph (TWF) that is to be stored at each site. A node called "external" or Ex is a basic entity involved in this algorithm that abstracts nonlocal portion of the global TWF graph. Ron Obermarck [15] in his paper gave a proof of correctness for this algorithm of distributed deadlock detection including an example of the algorithm in operation along with the performance characteristics of the algorithm.

The process for deadlock detection is as follows:

→ A particular site waits for deadlock-related information to be received from other sites.

→ As soon as the information is received by this site it then combines this information with its local TWF graph. And then detects all cycles and the cycle that does not contain the node Ex is broken.

→ If the node Ex is contained in the cycles  $Ex \rightarrow T_1 \rightarrow T_2 \rightarrow Ex$ , the particular site sends them in a string form  $Ex, T_1, T_2, Ex$  to all the other sites.

In this algorithm message traffic is reduced and a string say  $Ex, T_1, T_2, T_3, Ex$  is sent to other sites only if  $T_1$  is higher than  $T_3$  in the lexical ordering [5].

iv. *Chandy- Misra- Haas algorithm*- a concept of special message called probe is used in this algorithm. A probe is defined by a set ( i, j, k). It means that in this set

a deadlock detection is initiated for process  $P_i$  that is being sent by home site of process  $P_j$  to the home site of process  $P_k$ . The probe message travels along the edges of the global TWF graph and we say deadlock is detected when this message returns to its initiating process.

Chandy, Misra and Haas [16] in their paper presented a distributed deadlock models and also said that no false deadlocks are reported.

**4.3.3.3. Hierarchical deadlock detection algorithm-** in this algorithm all the sites are arranged in hierarchical manner. A particular site is responsible for detecting deadlocks that involves only its children sites [5].

*i. Menasce-Muntz algorithm-* in this algorithm all the resource controllers are arranged in the form of a tree. In the hierarchy the controllers that manage resources are locate at the lowest level, and are referred to as leaf controllers. These leaf controllers also perform the task of deadlock detection and also maintains the part of the global TWF graph that is concerned with resource allocation at the leaf controller. A TWF graph is maintained by a non leaf controller that spans only its children controller and detects deadlocks that involves its own leaf controllers. Any change in the TWF graph of a controller as a result of resource allocation, wait, or release causes an appropriate change in its parent controller [5]. Now this parent controller makes appropriate changes in the TWF graph and searches for any cycle. And if necessary these changes are sent up in the hierarchy.

*ii. Ho-Ramamoorthy algorithm-* in Ho and Ramamoorthy 's hierarchical algorithm the sites are grouped to form several clusters. A site is periodically chosen as the central control site and it then dynamically chooses control site for each cluster. This central site can request other control sites for their inter cluster transaction status information and wait-for relations. With the help of this process at a particular control site, status tables of all the sites is collected in its cluster and then one-phase deadlock detection algorithm is applied to detect all deadlocks involving only intra cluster transactions. This site then sends the inter cluster transactions status information and wait-for relations to the central site. This helps the central site to construct the system state graph, and thus perform checks for the presence of any cycle. Therefore, it can be summed up and said that the control site detects deadlocks that are present in its own cluster and central site detects all inter cluster deadlocks.

Edgar Knapp [17], in his paper presented basic principles on which distributed deadlock detection schemes are based. He also said that these principles provide a method to develop distributed algorithms. In his paper he discussed number of algorithms and also their respective complexities.

Out of all the deadlock detection algorithms discussed so far, for Multi User Virtual Drawing Board the distributed approach is the best to be used in order to achieve an extensible distributed framework. All the distributed deadlock algorithm aims to detect cycles not only at a particular site but spans several sites in the system. Now the way this aim is achieved by distributed deadlock detection algorithm differs according to the method used that is the type of distributed algorithm used (Goldman's, Obermarck's, Isloor Marsland's etc.)

Therefore, it depends on the type of distributed deadlock detection algorithm we are using that have its associated advantages and disadvantages that makes an application more reliable and secure.

**4.4 Concurrency management-** several processes in a distributed system are said to be concurrent if they perform their task at the same time. Now this concurrency may give rise to several problems like 'inconsistent update' problem and 'inconsistent retrievals' problem [6]. Inconsistent retrieval refers to the situation when a particular transaction reads some data objects of a database before another transaction has completed some modifications on those data objects [9]. Inconsistent update refers to the situation when on a common set of data objects many transactions perform read and write that leads to the inconsistency in the database [9].

Various concurrency control algorithms are:

**4.4.1. Locking-** is most widely used algorithm for concurrency control. In locking method a transaction locks a data item before actually accessing it. Now this transaction can access this data item any number of times. And no other transaction can have access to this data unless it is released by the transaction that acquired a lock over it[2].

**4.4.2. Optimistic concurrency control-** was given by Kung and Robinson [8]. This method allows the transactions to continue until the end of first phase. But in the second phase the before a transaction is committed, the transaction is first validated to check any inconsistency caused by any other transaction since it is started. The transaction is committed only if it is found valid otherwise it is aborted.

**4.4.3. Timestamps-** in this method a transaction is assigned a unique timestamp at the time it is started. Every data is assigned two timestamps that is *read timestamp and write timestamp*. Whenever a transaction wants to access a data then the data item's read timestamp or write timestamp is updated according to the transactions timestamp depending on the type of access [2].

Ricart and Agrawala in 1981 [18] said that if a process wants to enter the critical section it should ask the other processes to give it the permission to enter into the

critical section. Therefore, it waits for the appropriate permissions for the same. For this reason Ricart-Agrawala proposed the timestamp mechanism in order to associate a timestamp for each request. Since these timestamps are ordered therefore, conflict problem is resolved. The request having the lowest timestamp amongst all the conflicting requests gets the highest priority and by this conflicts are resolved.

Michel Raynal [19], in his paper discussed the principles from which the distributed mutual exclusion algorithms are designed. These principles include permission-based and token-based principles.

Permission-based algorithms include the one that was given by Ricart and Agrawala that is based on timestamps.

In token-based method, tokens are used to grant permissions to the processes that want to enter into the critical section and these processes waits until the token arrives.

Kerry Raymond [20] in his paper proposed a tree-based algorithm for distributed mutual exclusion. According to this method in a system there are  $N$  nodes that communicate to each other by means of passing messages. In this algorithm a spanning tree of the system is used. The number of messages exchanged per critical section depends on the topology of the tree. Each node in this tree has the capability of storing information about their immediate neighbour rather than about all nodes that are not in the spanning tree. The nodes that fail can recover all the necessary information from their neighbours. This algorithm does not require the use of sequence numbers because it operates correctly without the use of it also.

Suzuki and Kasami [21] in their paper gave a distributed mutual exclusion algorithm. In this algorithm mutual

exclusion is carried out among  $N$  nodes in a system and it requires  $N$  message exchanges for each mutual exclusion invocation. With this algorithm the delay that was caused to invoke mutual exclusion is much smaller in comparison to the algorithm for mutual exclusion given by Ricart and Agrawala in which  $2^*(N-1)$  message exchanges are required per invocation. The drawback of this algorithm given by Suzuki and Kasami is that it uses sequence number concept and these sequence numbers contained in the messages are unbounded but in this paper a method is given to resolve this problem by slightly increasing the number of message exchanges.

For Multi User Virtual Drawing Board concurrency management can be done by applying serialization protocol in order to perform simultaneous updates.

**4.5 Consensus-** a consensus problem is a situation where each and every processor broadcasts its initial value to all the other processors in the system [9]. Now this initial value may be different for all the processors. So a protocol for reaching consensus is required and this protocol should meet the following conditions:

→ **Agreement:** all the processors that are non-faulty must agree on the same single value.

→ **Validity:** now if the non-faulty processors have an initial value of  $v$  then the common value on which the processors must agree should be  $v$ .

Michael J. Fischer *et.al* [10] in his paper discussed about consensus problem. In his paper he said that consensus problem involves an asynchronous system of processes and some of these processes may be unreliable. He also discussed about “Byzantine Generals” problem.

For Multi User Virtual Drawing Board consensus problem can be solved by using a primary replication server which can be used to continuously send updates and check for acknowledgement.

Table 2: Distinction between three Agreement problems in Distributed Systems

	Byzantine Agreement	Consensus	Interactive Consistency
1. Initiator of the value	One processor initiates the value	All processors	All processors
2. Final agreement	Single value	Single value	A vector of values
3. solution of the problem	Solved by solution to the consensus problem	Solved by solution to the interactive consistency problem.	Byzantine agreement problem is a primitive to solve this problem.

**4.6 Security in distributed systems-** since in a distributed system, nodes are located at various locations. So a major concern in a distributed system is the security of an application. Robert Cole [22] in his paper gave a model of security in distributed systems. And also certain issues on the use of this security model were given.

There are various threats that are needed to be addressed in a distributed system [22] like information disclosure,

use of resources by unauthorised means, repudiation of information flow, denial of service, information contamination, misuse of resources.

A model of security in a distributed system should be such that it should cover all the security threats and this model should well adapt to suit different security policies. He also defined the concept of security information that is always generated by a specific

security services in response of an authenticated requests. And the main purpose of any security service is to use security information appropriately in a distributed system [22].

Dan M. Nessett [23] in his paper gave certain factors that affect the security of distributed systems including factor like node evaluation levels and network topology.

Rob Dobry [24] in his paper discussed various aspects of security required in distributed systems. The various aspects include cryptography, data confidentiality, and data integrity. He said in order to develop a secure distributed system, a system should not only make use of traditional computer security concepts but must also utilize communication security concepts.

Alan H. Karp [25] in his paper gave three components of the system architecture to make it easier to manage and monitor distributed systems. The work he presented was based on certain assumptions like large number of machines and users, dynamic, heterogeneous, hostile, different environments. The three components used in this paper were separate granting of rights from access control, mediate between applications and user and lastly using a proxy for remote users.

Security in MUVDB will be achieved in a way that no user should be able to delay network traffic or cause denial of service and during consensus a user is not able to vote multiple times. We handle the first issue by diagnosing abrupt traffic generating nodes and further eliminating them from our multicast network. The second issue is handled by authenticating users with user database and allowing only authentic users to participate in the application. We also make voting during consensus as an idempotent operation.

## 5. Conclusion

A distributed system consists of various computers located at different sites. Each of these computers have their own local memory and processors. All these computers are connected by means of a communication network. As seen through this paper that there are several issues related to the distributed system and all these issues have their respective advantages and disadvantages. A message passing system should be such that it should deliver the message from sender to receiver without leading to any type of error. Clocks must be synchronized to avoid the problem of clock skew and clock drift. And for a distributed system it is appropriate to use vector clocks for global ordering of events. As already discussed that amongst the three deadlock handling techniques deadlock detection is most advantageous in comparison to deadlock avoidance and deadlock prevention and there are various deadlock detection algorithms that can be applied depending upon the need of the system. In case of concurrency

management through this paper it can be concluded that timestamps technique for concurrency control is better in comparison to the other two techniques discussed in the paper. As far as consensus problem is concerned it is clear through the table for the three agreement problems that all these agreement problems are complementary to each other. The security of the distributed system is the key aspect while developing any distributed system. Security includes various threats that should be addressed whenever distributed systems are taken into account.

## 6. Future Scope

The idea presented in this paper for an extensible distributed framework can be implemented to develop a Multi User Virtual Drawing Board. A Multi User Virtual Drawing Board can support multiple users at a same time to solve the problem of collaborative designing. The idea behind this Multi User Virtual Drawing Board is to provide users with a virtual drawing board and all changes made by each user are reflected in all the virtual boards. For this various concepts of distributed system can be used like message passing, sending updates, making consensus for agreement in case of conflicts, concurrency management etc. Interface for such an application can be designed using swings and a message passing model using sockets and threads.

## References

- [1] Latha CA, Dr. Shashidhara HL, "Clock Synchronization in Distributed Systems", *2010 5<sup>th</sup> International Conference on Industrial and Information Systems, ICIIIS 2010*, Jul 29-Aug 01, India.
- [2] Pradeep K. Sinha "Distributed Systems" concepts and design, *Prentice-Hall, India*.
- [3] Kenneth P. Birman, "The Process Group Approach to Reliable Distributed Computing", *communication of the ACM* December 1993/vol.36, no.12.
- [4] Hagit Attiya, Amotz Bar-Noy and Danny Dolev, "Sharing Memory Robustly in Message-Passing Systems", *journal of the association for computing machinery*, vol 42, no. 1, January 1995, pp 124-142.
- [5] Mukesh Singhal, "Deadlock Detection in Distributed System", *Ohio State University, November 1989, IEEE*.
- [6] George Coulouris, "Distributed Systems concepts and design", *fourth edition, pearson education*.
- [7] Leslie Lamport, "Time, Clocks, and the Ordering of Events in a Distributed System", *communication of the ACM*, july 1978, vol 21, no. 7.
- [8] Kung and Robinson, "On Optimistic Methods for Concurrency Control", *ACM transactions on database systems*, vol. 6, no. 2, pp. 213-226 (1981).
- [9] Mukesh Singhal, Niranjana G. Shivaratri, "Advanced Concepts in Operating Systems", *Tata McGraw-Hill edition*.



- [10] Michael J. Fischer, Nancy A. Lynch and Michael S. Paterson, "Impossibility of Distributed Consensus with One Faulty Process", *Journal of the Association for Computing Machinery*, vol. 32, no. 2.
- [11] Kasim Sinan Yildirim and Aylin Kantarci, "Clock Synchronization in Distributed Systems", Computer Engineering Department, Ege University. *The Turkish Scientific and Technical Research Council (TUBITAK)*.
- [12] R.Baldoni, M.Raynal, "Fundamentals of Distributed Computing: A Practical Tour of Vector Clock Systems".
- [13] Birman K and Joseph T, "Reliable Communication in the presence of failures". *ACM Transactions on Computer Systems*, 5(1): 47-76, 1987.
- [14] Li-Hsing Yen and Ting-Lu Huang, "Resetting Vector Clocks in Distributed Systems", *Journal of Parallel and Distributed Computing* 43, 15-20 (1997, Article No.PC971330).
- [15] Ron Obermarck, "Distributed Deadlock Detection Algorithm", *ACM Transactions on Database Systems*, vol.7, No.2, June 1982, 187-208.
- [16] K. Mani Chandy, Jayadev Misra and Laura M. Haas, "Distributed Deadlock Detection", *ACM Transactions on Computer Systems*, Vol. 1, No. 2, May 1983, 144-156.
- [17] Edgar Knapp, "Deadlock Detection in Distributed Databases", *ACM Computing Surveys*, Vol. 19, No. 4, December 1987.
- [18] Ricart G. Agrawala A.K, "An Optimal Algorithm for Mutual Exclusion in Computer Networks", *Comm.ACM*, Vol. 24, 1, (1981), pp.9-17.
- [19] Michel Raynal, "A Simple Taxonomy for Distributed Mutual Exclusion Algorithms".
- [20] Kerry Raymond, "A Tree-Based Algorithm for Distributed Mutual Exclusion", *ACM Transactions on Computer Systems*, Vol. 7, No. 1, February 1989, pages 61-77.
- [21] Ichiro Suzuki and Tadao Kasami, "A Distributed Mutual Exclusion Algorithm", *ACM Transaction on Computer Systems*, Vol. 3, No.4, November 1985, pages 344-349.
- [22] Robert Cole, "Security for Distributed Systems", *Hewlett-Packard Laboratories*.
- [23] Dan .M. Nessett, "Factors Affecting Distributed System Security", *IEEE Transactions on Software Engineering*, Vol. SE-13, No. 2, February 1987.
- [24] Rob Dobry and Mary D. Schanken, "Security Concerns for Distributed Systems", *IEEE 1994*.
- [25] Alan H. Karp and Kevin Smathers, "Three Design Patterns for Secure Distributed Systems", *Hewlett-Packard Company 2003*.



**Shilpi Gupta** is working as an assistant professor in Department of Computer Science and Engineering of Amity School of Engineering & Technology, Amity University, Noida. She has 05 years of experience in the field of Academics and is actively involved in research & development activities. She has received her B.Tech degree in 2006. She is M.Tech Gold medalist in the stream of Computer Science and Engineering from Jaypee Institute of Information Technology, Noida. Her area of interest includes Software Engineering, Artificial Intelligence, Soft Computing, Cognitive Informatics, Affective Computing. She has successfully published national and international research papers.



**Aneesha Sharma** is a student who has received her B.Tech degree in Computer Science and Engineering in 2010 with first division from Amity School of Engineering and Technology, Amity University, Noida, Uttar Pradesh, India. Currently, the author is pursuing M.Tech in Computer Science and Engineering from Amity School of Engineering and Technology, Amity University, Noida, Uttar Pradesh, India. The author is currently undergoing a dissertation period in the field of Distributed Systems.

# A SLA-Aware Scheduling Architecture in Grid System Using Learning Techniques

Seyedeh Yasaman Rashida<sup>1</sup>, Amir Masoud Rahmani<sup>2</sup>

<sup>1</sup>Department of Computer Engineering, Shirgah Branch, Islamic Azad University  
Shirgah, Mazandaran, Iran

<sup>2</sup>Department of Computer Engineering, Islamic Azad University, Science and Research Branch  
Tehran, Tehran, Iran

## Abstract

In the Grid environment, the relationship between a customer and a service provider should be clearly defined. The responsibility of each partner can be stated in the so-called Service Level Agreement (SLA). A SLA is a formal contract between end-user and system to guarantee that customers' service quality expectation can be achieved. In recent years, extensive research has been conducted in the area of SLA for utilizing computing systems and also, various SLA-based scheduling are proposed but the number of resources and tasks to be scheduled is usually variable and dynamic in nature. Most of proposed algorithms don't have flexibility in all situations, because every scheduling algorithm cannot improve all grid factors like resource utilization, load balancing, etc and cannot notice all parameters at the moment. In this paper, we propose SLA aware scheduling architecture which uses learning techniques for selecting best way to schedule resources in different situations. The proposed model causes increasing user satisfaction, number of completed tasks and system utilization and resource load balancing. At the end, we formulize relation between number of completed tasks and system utilization.

**Keywords:** SLA, Scheduling, Grid Computing, Learning Technique, Load Balancing.

## 1. Introduction

Grid computing system is a collection of distributed heterogeneous computing resources available over a local or wide area network that appears to an end user or application as one large virtual computing system. Grid computing is to provide an unlimited power, collaboration, and information access to everyone connected to grid [1].

A schedule is defined as a function  $f: T \rightarrow R$  which maps every task  $T_i \in T$  on a resource  $R_j \in R$  that has attached to a queue  $Q_j$ . The goal of any schedule is to minimize the cost function such as scalability and lateness.

The grid scheduler has four phases, which consists of resource discovery, resource selection, job selection and job execution. A grid scheduler acts as an interface between the user and distributed resources. It hides the complexity of the computational grid from the grid user.

The main responsibility of a scheduler is selecting resources and scheduling tasks in such a way that the user and application constraints are satisfied, in terms of overall execution time and cost of the resources utilized. To achieve these goals, Service Level Agreement (SLA) can play a critical role. In general, SLAs are defined as an explicit statement of expectations and obligations in a business relationship between service providers and customers. SLAs specify priori negotiated resource requirements, the quality of service (QoS), and costs.

Most research applies one or two scheduling algorithms to achieve their goals such as maximize number of completed jobs, system utilization, etc. But it is important to notice that each scheduling algorithm can improve some of the expected factors. In order to dynamic grid environment, it is possible to confront critical situation which applied proposed scheduling cannot achieve the whole goal. In this paper, we propose a SLA-aware scheduling scheme which uses learning techniques to select best way of scheduling for achieving system goals in variable situations.

The rest of this paper is organized as following. In section 2, we discuss related work. Section 3 describes the proposed scheduling architecture. In section 4, we obtain relationship between number of completed jobs and utilization. Section 5 gives the concluding remarks.

## 2. Related Work

Distributed resource allocation is one of the most challenging problems in resource management field. This problem has attracted a lot of attention from the research community in the last few years. In the following we provide a review of some relevant prior work.

Bin Zeng et al. [5] propose a negotiation based model, where adaptive learning agents, representing individual resources and tasks, co-operate among themselves to help achieving a near optimal schedule. N.Malarvizhi and V.Rhymend Uthariaraj [6] describe a scalable grid-architecture involving a Grid Resource Manager,

assuming the role of a resource broker to select computational resources based on job requirements and the capacity of grid resources, so as to minimize the time to process each application along with transmission time associated with it. D. P. Spooner et al. [7] develop a multi-tiered scheduling architecture

(TITAN) that uses a performance prediction system (PACE), along with brokers that are involved in distribution of jobs in the grid, to meet deadlines and significantly increase the efficiency of resource utilization.

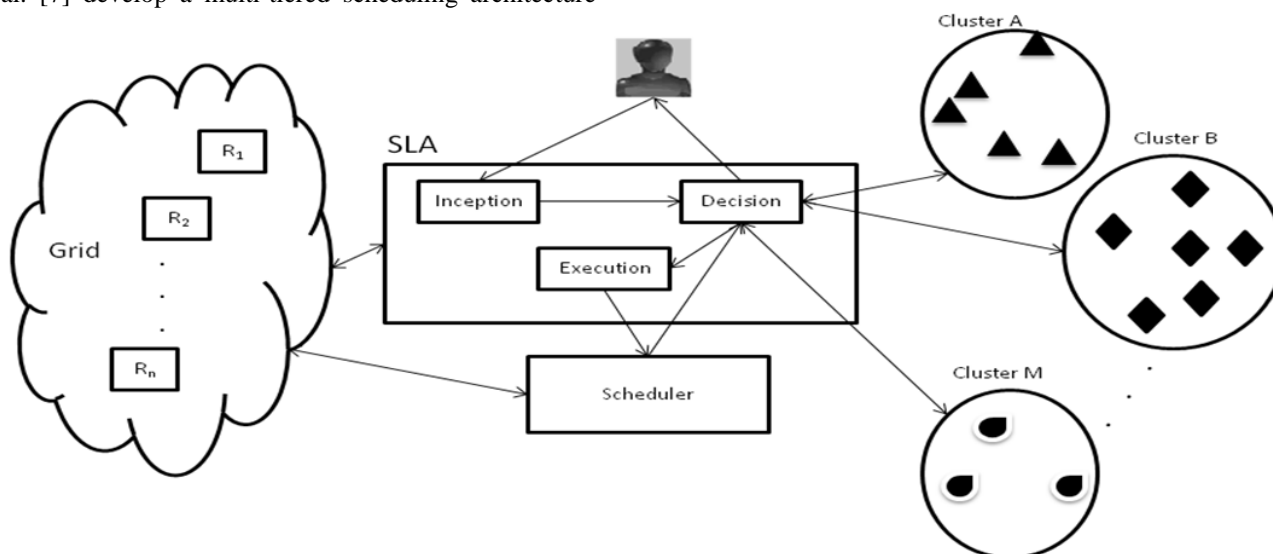


Fig.1. SLA-aware Scheduling Architecture

The paper [8] presents a novel load balancing approach in a heterogeneous distributed environment. The scheduler takes into account the threshold value, based on the ratio of service rates, along with the queue length to determine whether it is beneficial to migrate a given local task to another node in the system or not. Markov process model is used to describe the behavior of the heterogeneous distributed system under the proposed policies. Kumar also proposes a Load balancing algorithm for fair scheduling, and compares it to other scheduling schemes such as the Earliest Deadline First, Simple Fair Task order, Adjusted Fair Task Order and Max Min Fair Scheduling for a computational grid. It addresses the fairness issues by using mean waiting time. It scheduled the task by using fair completion time and rescheduled by using mean waiting time of each task to obtain load balance. This algorithm scheme tries to provide optimal solution so that it reduces the execution time and expected price for the execution of all the jobs in the grid system is minimized [27]. In [30], Anandharajan and Bhagyaveni propose to find the best EFFICIENT cloud resource by Co-operative Power aware Scheduled Load Balancing solution to the Cloud load balancing problem. The algorithm developed combines the inherent efficiency of the centralized approach, energy efficient and the fault-tolerant nature of the distributed environment like Cloud. Shahu Chatrapati et al. [28] propose *Competitive Equilibrium Scheme* (CES) that simultaneously minimizes mean response time of all jobs, and the response time of each job individually. Ruay-Shiung Chang et al. [9] propose an Adaptive Scoring Job Scheduling algorithm (ASJS) for a distributed grid environment to reduce the completion

time of submitted jobs, by assigning jobs to resources after looking into recent scheduling history of every available resource and then choosing the most optimal one.

Computing intensive jobs and data intensive jobs handled differently, and local and global updates are used to obtain the most recent status of grid resources to schedule jobs more effectively in real time. System Model Syed Nasir Mehmood Shah et al. [10] propose an algorithm for CPU scheduling of a modern multiprogramming operating system, design and development of new CPU scheduling algorithms (the Hybrid Scheduling Algorithm and the Dual Queue Scheduling Algorithm) with a view to minimize overall task schedule. The following paper extends this prioritized round robin heuristic from a single system multiprogramming environment, onto a multi-processor distributed architecture. As each scheduling strategy optimizes some of performance parameters such as making span, resource utilization, response time, workload balancing, service time, reliability, fairness deviation and throughput, we propose a SLA-aware scheduling model to achieve four important parameters such as resource utilization, response time, workload balancing and throughput.

In [26], Murugesan and Chellappan introduce a new resource allocation model with multiple load originating processors as an economic model. Solutions for an optimal allocation of fraction of loads to nodes obtained to minimize the cost of the grid users via linear programming approach. It is found that the resource allocation model can effectively allocate workloads to proper resources.

In [25], presents a clustering technique for gene expression data which can also handle incremental data. It is called GenClus and designed based on density based approach. Experimental results show the efficiency of GenClus in detecting quality clusters over gene expression data. Our approach improves the cluster quality by identifying sub-clusters within big clusters.

### 3. Proposed Scheduling Architecture

We represent a SLA aware scheduling architecture (Given at fig. 1) with making decision ability to achieve grid goals such as increasing resource utilization, number of completed jobs, percentage of user satisfaction, load balancing, etc. In the proposed model, SLA encompasses three parts– inception, decision and execution–that make decision based on request status and system condition what scheduler policy would be best way to achieve grid goals.

As shown fig. 2, requests and system status as input would be given to SLA. SLA selects some operations as action to operate on inputs. Based on action which be done on inputs, grid system status would be changed, on the other hands, grid goals would be changed. So it is important to apply best action for achieving grid goals.

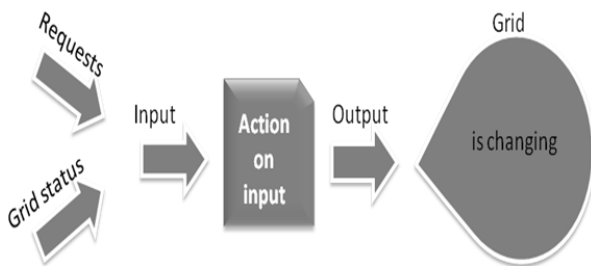


Fig.2. General representation of proposed procedure

#### 3.1. How To Specify Inputs

In this section, we discuss the way for specifying requests and system status as inputs of our model.

##### 3.1.1 Request Properties

SLA requires information in both requests (jobs) and system status, in order to making the decision to do a function (action) mapping inputs–requests and system status–to desired outputs. Since Deadline, service time, priority of jobs and system workload are very important factors for SLA to recognize current status and make decision what to act for producing sufficient output, we use these parameters to obtain required information and define a job as follow:

$$J = \langle R, Q, D, S, P \rangle$$

Each job will have some requirements as resource,  $R$ , quality of service of resource,  $Q$ , job deadline,  $D$ , job service time,  $S$ , job priority,  $P$ .

Jobs prioritize based on applied action. For example, privileged program priority would be more than batch job priority. Job deadline is made on two parts, service time,  $S$ , and laxity,  $L$ , as shown in eq. 1. *Service time* is the time a job takes to finish executing on resources. *Laxity* is the time a job holds resources with no using, as shown in fig. 3.

$$D = S + L \quad (1)$$

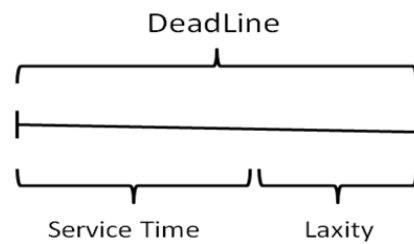


Fig. 3. Division of Deadline

Since there are numerous submitted jobs at the moment, selecting each job as discrete input and surveying them discretely will increase computing overhead. So, we suppose a time slot that jobs are surveyed at the start of time slot. Based on system workload, time slot fluctuates in time. It means if system workload is high, then time slot would grow until system workload decreased and vice versa.

For specifying jobs status as part of input, it requires to calculate mean of request deadline,  $\mu_d$ , and mean of request service time,  $\mu_s$  and also mean of request priority,  $\mu_p$ , at start of time slot, as shown in eq. (2)-(4). In our model, jobs status would be defined by these three parameters.

Table 1. Characteristics of the proposed system

n	number of submitted jobs at the start of time slot
$P_i$	priority of $i$ th request
$D_i$	deadline of $i$ th request
$S_i$	service time of $i$ th request
$\mu_d$	mean of request deadline
$\mu_s$	mean of request service time
$\mu_p$	mean of request priority

$$\mu_d = \frac{\sum_{i=1}^n P_i \times D_i}{n} \quad (2)$$

$$\mu_s = \frac{\sum_{i=1}^n P_i \times S_i}{n} \quad (3)$$

$$\mu_P = \frac{\sum_{i=1}^n P_i}{n} \quad (4)$$

### 3.1.2 Specifying System Status

Inputs are job and system status. In previous section, we explained how to specify job status and in this section, we want to explain how to specify system status. For obtaining information about the system, there are some smart agents monitoring system status and alarming the status to the SLA. SLA notices obtained information about requests and system status as inputs and proceeds to set best action to do on the input so that desired output has been produced. In the next section, it will be described how to select best action regarding to inputs.

### 3.2 How To Choose Action

As explained later, based on inputs, SLA should choose a sufficient action to obtain desired output-resource utilization, number of completed jobs, percentage of user satisfaction, load balancing. It is difficult to map inputs to desired output. To solve the problem, we use clustering technique together with supervised learning. Cluster analysis or clustering is the process of grouping the objects into subsets so that the objects in subset are similar in some sense. Clustering is a method of unsupervised learning and a common technique for statistical data analysis used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics [29]. Fig. 4 represents the clustering process. In this section, we discuss algorithms and methods implemented for grouping similar inputs and generating sufficient actions.

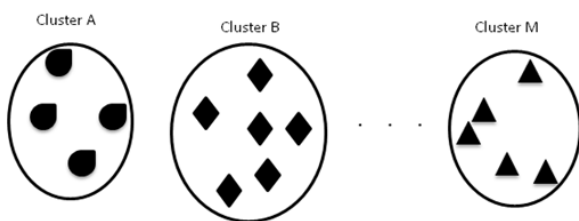


Fig. 4. representation of clusters

#### 3.2.1 Clustering Algorithm For Grouping Inputs

Cluster is the same template of items- Input, Action and Output- which have been registered in tuple with fields are listed eq. 5. The first four parameters are related to inputs and the others are related to system and actions. As shown on eq. 6, system status comprises two options - load balancing rate and resources utilization rate - that alarm by smart agents. And as shown in eq. 7, request status comprises two field- percentage of user satisfaction and percentage of completed jobs. Percentage of completed jobs is

computed easily but it needs to compute some parameters in order to obtain percentage of user satisfaction that it is shown in eq. 8. All parameters which exist in eq. 5-8, are listed on table 2.

$$\langle \mu_D, \mu_S, \mu_P, request - status, systemstatus_{current}, action_{related}, systemstatus_{next}, T_{Slot} \rangle \quad (5)$$

$$systemstatus = \langle load - balance, utilization rate \rangle \quad (6)$$

$$request - status = \langle percentage\ of\ satisfaction, percentage\ of\ completed\ jobs \rangle \quad (7)$$

$$P_{satisfy} = p_i \left( D_i - \left( \frac{T_{Sij}}{S_i^{min}} + \frac{L_i}{Bw_{ij}} T_{ij} \right) \right) + p'_i (1 - p_i) (D_i - T_c - T_N) \quad (8)$$

Table 2. Existing parameters in eq. 5-8

$T_{Slot}$	Determined time slot by system
$action_{related}$	Selected action for the input
$systemstatus_{next}$	System status after doing action on inputs
$status_{request}$	Status of requests which submitted to SLA
$systemstatus_{current}$	Current system status
$P_{satisfy}$	Percentage of user satisfaction
$D_i$	Deadline of job $i$
$T_{Sij}$	The $i$ -th job service time on the resource $j$
$S_i^{min}$	Speed of the slowest resource on which the job $i$ can be executed
$L_i$	The size of a given job $i$
$Bw_{ij}$	The bandwidth between $i$ -th job and the resource $j$ on which the job can be executed
$p'_i$	Probability that $i$ -th job submitted for the second negotiation
$p_i$	Probability that $i$ -th job submitted for the first negotiation
$T_{ij}$	Required time to transfer data from $i$ -th job to resource $j$
$T_c$	Completed job time
$T_N$	Required duration for the first negotiation

At first, there is no knowledge about system, on the other hands, no cluster exists. So selecting randomly an action to apply on requests called as inputs can cause undesirable results. For solving the problem, we early apply supervised technique. It means that grid system manger supposes some inputs which might happen in system and then manager makes decision which action or policy is conducted to goals. Supposed inputs and system status are classified on different clusters. Each cluster includes group of similar status, as shown in fig. 4.

As shown in fig. 5, we describe only six statuses with proposed action across all statuses which grid manager can suppose:

*Status 1:* Jobs deadline are too short and they have high priority. It means jobs should do as fast as possible with low rejection rate.

*Action 1:* Scheduler should use a quick searching algorithm to find sufficient resources with matching requests like hill climbing. It is better to find resources nearby request because transmitting jobs to remote resources would spend time. Even if there is no idle local resource, it is better to find resource by which young request,  $J_i$ , with long deadline is executed. The new request is sent to the resource and  $J_i$  is sent to remote idle resource.

*Status 2:* Jobs deadline are too short but no have high priority and system workload is high. And there is no scarcity of resources. It means jobs should be done fast in order to more accepted jobs. But it is never forget that load system is high.

*Action 2:* Scheduler should use a quick searching algorithm to find sufficient resources with matching requests like hill climbing. It is better to find resources nearby request Because of transmitting jobs to remote resources would spend time. Also, Time slot,  $T_{slot}$ , should grow because of heavy system load, it causes that fewer number of job would be submitted so system load would decline.

*Status 3:* Jobs deadline are normal but there is no load balancing.

*Action 3:* Scheduler should use a searching algorithm to provide system load balancing like BACO (Balance ant colony optimization).

*Status 4:* Jobs deadline is normal but system workload is high and there is scarcity of resources.

*Action 4:* Scheduler can use an Economic heuristic algorithm to find sufficient resources with matching requests like Game theory.

*Status 5:* Jobs deadline are too short. System does not have heavy system workload.

*Action 5:* Scheduler should use hill climbing algorithm to search resources nearby request.

*Status 6:* Jobs deadline are normal and there is load balancing.

*Action 6:* Scheduler use PSO algorithm to search resources.

As illustrated above, we only use four scheduling algorithms, PSO (partial swarm optimization), BACO (balanced ant colony optimization), Hill climbing and Economic based heuristic like game theory. Each algorithm can improve some grid factors. For example, BACO is capable of achieving system load balance better than other job scheduling algorithms and also economic heuristic deals with matching jobs to available resources in economical way such that resource provider and consumer get sufficient incentive to stay and play in competitive market [2, 3].

Each upper status makes a distinct cluster based on input, action and output. Input and action will be registered in related cluster but output will be registered after executing action on input and observing the result on grid.

Anyway, after requests submit to SLA, the new observation should be lied on clusters. For the purpose of grouping similar sets of inputs, we apply k-means clustering algorithm. The registered information on cluster is shown in eq. (5)-(7).

K-means is a clustering algorithm that, given an initial set of k means, assigns each observation to a cluster with the closest mean. It then calculates new means to be centers of observations in the clusters and stops when the assignments no longer change [22]. A cluster center is a newly generated input for a group of requests.

A frequent problem in k-means algorithm is the estimation of the number k. Two implemented approaches are explained as follow [23]:

1. *Rule-of-Thumb* is a simple but very effective method in which k is set to  $\sqrt{N/2}$ , where N is the number of entities.

2. *Hartigan's Index* is an internal index introduced in [24]. Let  $W(k)$  represents the sum of squared distances between cluster members and cluster center for k clusters. When grouping n items, the optimal number k is chosen so that the relative change of  $W(k)$  multiplied with the correction index  $\gamma(k) = n - k - 1$  does not significantly change for k + 1,

$$H(k) = \gamma(k) \frac{W(k) - W(k + 1)}{W(k + 1)} < 10$$

The threshold 10 shown in Hartigan's index is also used in our model. It is "a crude rule of thumb" suggested by Hartigan [24].

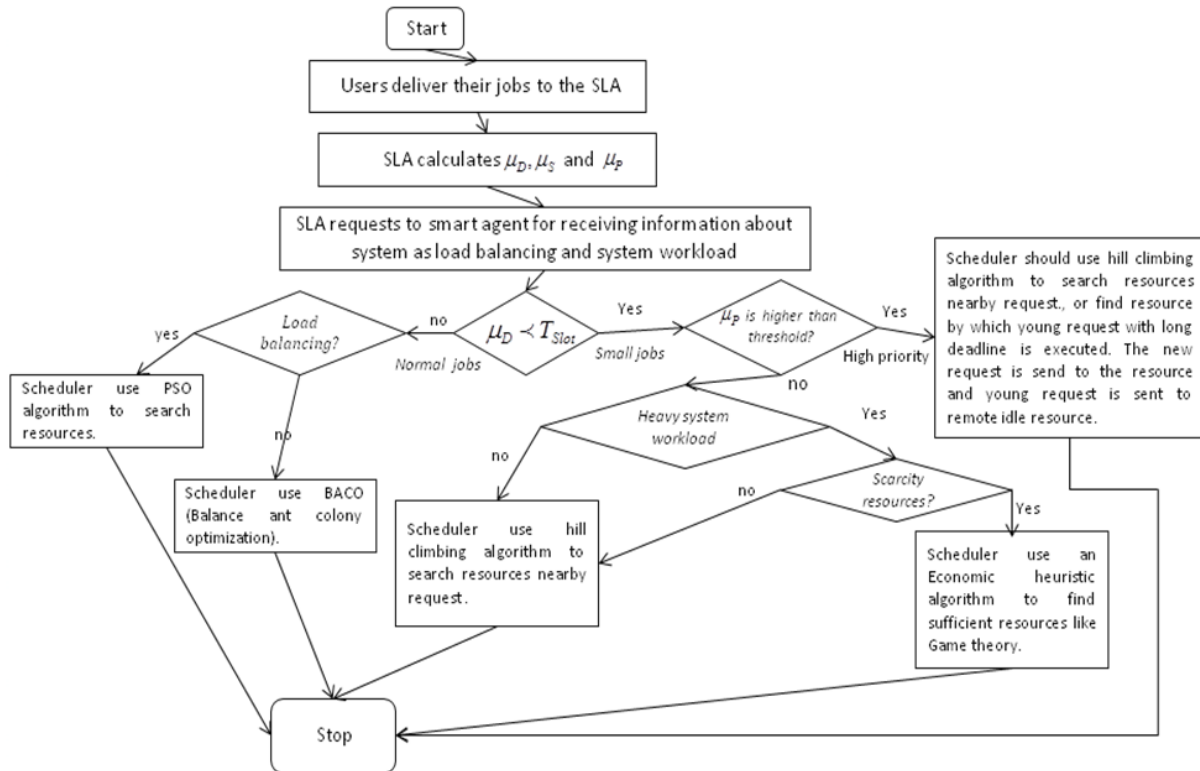


Fig.5. Describing six proposed initial statuses by system manager

### 3.2.2 Computing Distance Between Input Template

In order to utilize clustering algorithm, the measure of distance between two clustering items must be defined, as well as the distance between a clustering item and a cluster center. As already mentioned in the previous section, an item for clustering as called *Template* generally is a set of inputs, actions and outputs, while a cluster center is a master template for the given group of inputs.

Since having knowledge about the content of requests and about system status, we can exactly reconstruct templates. Therefore, both the distance between two clustering items as well as between an item and a cluster center can be reduced to computing distances between two templates. For this purpose, we introduce the n-tuple representation of decision. A decision consists of input, action and output parameters.

We distinguish between the structure of a decision, the list of input, action and output parameters with their names, and the values of a template, a list of numerical, boolean and string values of clustering items attributes. We introduce the n-tuple representation of a decision, where first  $n - 1$  elements contain values of the template, while the last element contains the template structure. By using such a representation, we can define the distance between two templates as an n-tuple, where first  $n - 1$  elements contain the differences between the two values of each of the parameters, while the final element contains a value

representing the difference between the structures of the templates.

To formalize, we observe two decision templates  $T_1$  and  $T_2$  defined by their values  $\{\alpha, \beta, \gamma, \dots\}$  and the template structure  $\tau$ .

$$T_1 = (\alpha_1, \beta_1, \gamma_1, \dots, \tau_1)$$

$$T_2 = (\alpha_2, \beta_2, \gamma_2, \dots, \tau_2)$$

The n-tuple  $D_{T_1, T_2}$  representing the distance between two clustering templates is defined as eq.(9) [23].

$$D_{T_1, T_2} = (f(\alpha_1, \alpha_2), f(\beta_1, \beta_2), \dots, F(\tau_1, \tau_2)) \quad (9)$$

The result of the function  $f$  for calculating the difference between two template values  $\alpha_1$  and  $\alpha_2$  depends on the type of its arguments and is defined as eq. (10) [23].

$$f(\alpha_1, \alpha_2) = \begin{cases} |\alpha_1 - \alpha_2|, & \text{if } \alpha_1 \text{ and } \alpha_2 \text{ are numerical} \\ 0, & \text{else if } \alpha_1 \text{ and } \alpha_2 \text{ are not} \\ & \text{numerical and } \alpha_1 = \alpha_2 \\ 1, & \text{else if } \alpha_1 \text{ and } \alpha_2 \text{ are not} \\ & \text{numerical and } \alpha_1 \neq \alpha_2 \end{cases} \quad (10)$$

The distance between the structures of clustering templates is expressed as a number of differences between properties of templates. This value is calculated by iterating through all parameters contained by at least one of the cluster templates, calculating the

distance between the templates with respect to the parameters. We define the distance function  $F$  calculating the difference between structures  $\tau_1$  and  $\tau_2$  of two templates  $T_1$  and  $T_2$  as shown in eq. (11) [23].

$$F(\tau_1, \tau_2) = \sum_{p \in T_1 \cup T_2} d_p(T_1, T_2) \quad (11)$$

where the distance between two parameters of two cluster templates with respect to its properties is defined as shown eq. (12).

$$d_p(T_1, T_2) = \begin{cases} 0, & \text{if name and metric of } p \text{ are} \\ & \text{the same in } T_1 \text{ and } T_2 \\ 1, & \text{else if } T_1 \text{ and } T_2 \text{ does not contain } p \\ 1, & \text{else if only name or metric of } p \\ & \text{differs in } T_1 \text{ and } T_2 \\ 2, & \text{else if both name and metric} \\ & \text{of } p \text{ differs in } T_1 \text{ and } T_2 \end{cases} \quad (12)$$

After the result tuple has been calculated, it can be used to generate a single numerical value representing the distance between the two clustering templates. In order to do so, the result tuple must be normalized beforehand so that the tuple elements can be mutually comparable. Normalization is executed on each of the  $n$  tuple elements separately, where a value of an element is divided by a range of possible values for the element (maximum value minus the minimum value). Then, the final value representing the distance between clustering items can be computed by a simple function. Note, in this paper, we discuss only the final element of the distance tuple, the difference between structures of two SLA templates. We plan to consider the values of SLA templates in our future work.

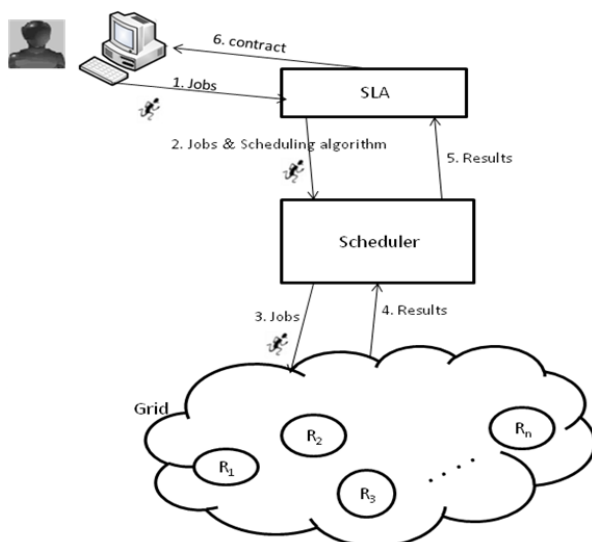


Fig. 6. Overview of the proposed model

However, SLA makes decision to choose sufficient action based on these clusters. If a status is found to which there is no matching cluster, SLA should make a

new cluster. SLA selects clusters which their inputs are nearby current input and based on these clusters, SLA guesses the sufficient action which is better in this status. Then new cluster would be created. So, as time pass, number of clusters will be changed.

It is important to note if there is a status which there is no algorithm to improve all of grid goals, SLA selects the algorithm that is able to improve user factors (number of completed jobs and user satisfaction) instead of system factors. In our model, user has high priority.

### 3.3 Glimpse Of The Proposed Model

As shown in fig. 6, requests submit to SLA. SLA surveys and analyses current grid status and received requests and based on input distinguishes related cluster for selecting sufficient scheduling policy to improve grid factors.

For example, it is possible that a status happens which SLA recommend user to change his job deadline otherwise job would be reject or other bad events happen. If user accepts, SLA's suggestion would be executed otherwise, user should resubmit his job. Described scenario of grid operation is illustrated in fig. 7.

### 3.4 SLA Infrastructure

In the proposed model, SLA encompasses three parts— inception, decision, execution— that make decision based on request status and system condition what scheduler policy would be best policy to achieve grid goals as shown in fig. 1.



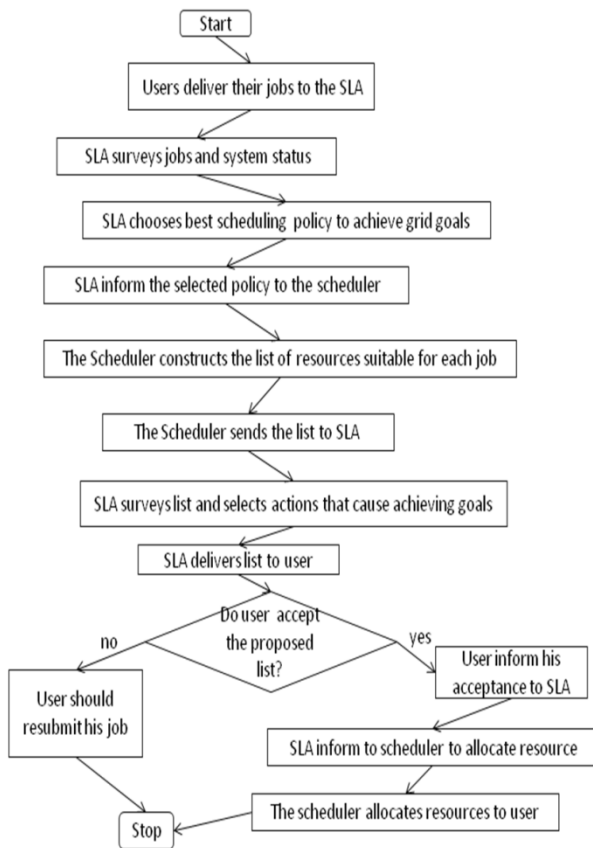


Fig.7. Describing scenario of grid operation

### 3.4.1 Inception Unit

When user submits his job to SLA, it is received by inception unit. The unit is responsible for surveying requests and system status as input and calculating mean of request deadline,  $\mu_d$ , and mean of request service time,  $\mu_s$  and also mean of request priority,  $\mu_p$ . And also communicates with smart agents to gain system status. Then sends obtained information to the decision unit.

### 3.4.2 Decision Unit

This unit is responsible to survey received result from the inception unit and determine the cluster which is related to current situation. If it distinguishes the cluster, then send the related action to the execution unit. But while no cluster match, the unit should create new cluster with current status as its input and choose its sufficient action using the approximately similar clusters but cannot specify its real output. The output would be specified after the action is done.

However, the unit chooses the action then sends it to the execution unit. After scheduler do the action and sends result to the decision unit, the unit surveys the result and selects best actions that best result would be provided for requests and system.

### 3.4.3 Execution Unit

It is responsible to send parameters to which scheduler requires executing selective action. Based on proposed action, scheduler finds resources then sends the result to the decision unit.

## 4. Relationship Between Number Of Completed Jobs and Utilization

In real world, there is no detailed mathematic relationship for most of phenomenon, so that specifying one phenomenon cause specifying the other one. For example, suppose that there is relation between company publicity and number of sell. The relation often is not precise, so that we can say if spend  $n$  dollar for publicity, company would sell  $q$  number of stuff. The relation that is between number of completed jobs and utilization is instance of this kind of relationship. First, draw transmittal diagram and then select best line which have minimum variance respect to spots in diagram. In this part, we intend to obtain relationship,  $\hat{r}$ , between the number of completed jobs,  $NCr$ , and system utilization,  $U$ , as illustrated in eq. (5) – (9). For specifying the relationship it needs  $n$  experiments which number of completed jobs,  $NCr_i$ , and system utility,  $U_i$ , in  $i$ -th experiment should measure in each experiment.

$$\mu_{NCr} = \sum_{i=1}^n \frac{NCr_i}{n} \quad (5)$$

$$\mu_U = \sum_{i=1}^n \frac{U_i}{n} \quad (6)$$

$$b = \frac{\sum_{i=1}^n NCr_i \times U_i - n \times \mu_{NCr} \times \mu_U}{\sum_{i=1}^n NCr_i^2 - n \times \mu_{NCr}^2} \quad (7)$$

$$a = \mu_U - b \times \mu_{NCr} \quad (8)$$

$$\hat{r} = a + b \times NCr \quad (9)$$

## 5. Conclusion

The following paper describes a novel SLA-aware model to schedule tasks efficiently in a grid environment. To apply best scheduling policy, this model uses clustering technique. Clustering is a method of unsupervised learning, and a common technique for statistical data analysis. Clustering is the assignment of a set of observations into subsets (called *clusters*) so that observations in the same cluster are similar in some sense. We put past scheduling experiments in disjoint clusters and in future, we use clusters to choose best scheduling policies.

## References

1. Y. Gao, H. Rong, J. Z. Huang, "Adaptive grid job scheduling with genetic algorithms", *J. Future Generation Computer Systems*, Vol. 21, 2005, pp. 151-161.
2. Y. S. Dai, X. L. Wang, "Optimal resource allocation on grid systems for maximizing service reliability using a genetic algorithm", *Reliability Engineering and System Safety*, Vol. 91, 2006, pp. 1071-1082.
3. A. V. Chandak, B. Sahoo, A. K. Turuk, "Heuristic Task Allocation Strategies for Computational Grid", *Int. J. Advanced Networking and Applications*, Vol. 2, 2011, pp.804-810.
4. D. D. H. Miriam, K. S. Easwarakumar, "A Double Min Min Algorithm for Task Metascheduler on Hypercubic P2P Grid systems", *International Journal of Computer Science Issues*, Vol. 7, No. 5, 2010, pp. 8-18.
5. B. Zeng, J. Wei, H. Liu, "Dynamic Grid Resource Scheduling Model Using Learning Agent", *IEEE International Conference on Networking, Architecture, and Storage (NAS'09)*, 2009, pp. 67-73.
6. N. Malarvizhi, V. R. Uthariaraj, "A Minimum Time To Release Job Scheduling Algorithm in Computational Grid Environment", *Proceedings of Fifth International Joint Conference on INC, IMC and IDC*, 2009, pp. 13-18.
7. D. P. Spooner, S. A. Jarvis, J. Cao, S. Saini, G. R. Nudd, "Local Grid Scheduling Techniques using Performance Prediction", *IEEE Proceedings of Computers and Digital techniques*, 2003, pp. 87-96.
8. S. Bansal, B. Kothari, C. Hota, "Dynamic Task-Scheduling in Grid Computing using Prioritized Round Robin Algorithm", *International Journal of Computer Science Issues*, Vol. 8, 2011, pp. 472-477.
9. M. Cochran, P. D. Witman, "Governance and Service Level Agreement Issues in A Cloud Computing Environment", *Journal of Information Technology Management*, Vol. 22, No. 2, 2011, pp. 41-55.
10. T. Altameem, "On the Design of Job Scheduling Strategy Using Agent Replication for Computational Grids", *IJCSNS International Journal of Computer Science and Network Security*, Vol. 11, No. 3, 2011, pp. 269-276.
11. J. Li, J. Peng, Z. Lei, W. Zhang, "An Energy-efficient Scheduling Approach Based on Private Clouds", *Journal of Information & Computational Science*, Vol. 8, No. 4, 2011, pp.716-724.
12. C. Ray, N. Guha, "Determination of Cost Model for Constraint based Query Optimization in Data Grids", *Proceedings of International Conference on Advances in Computer Science*, 2010, pp. 237-240.
13. L. M. Khanli, M. Etminan Far, A. Ghaffari, "Reliable Job Scheduler using RFOH in Grid Computing", *Journal of Emerging Trends in Computing and Information Sciences*, Vol. 1, No. 1, 2010, pp. 43-47.
14. M. Amoon, M. Mowafy, T. Altameem, "A Multiagent-Based System for Scheduling Jobs in Computational Grids", *ICGST- AIML journal*, Vol. 9, 2009, pp. 19-27.
15. M. Hovestadt, "Operation of an SLA-aware Grid Fabric", *Journal of Computer Science*, Vol. 2, No. 6, 2006, pp. 550-557.
16. R. J. S. Raj, V. Vasudevan, "Beyond Simulated Annealing in Grid Scheduling", *International Journal on Computer Science and Engineering (IJCSSE)*, Vol. 3, No. 3, 2011, pp. 1312-1318.
17. R. M. R. Kovvur, S. Ramachandram, V. Kadappa, A. Govardhan, "A Reliable Distributed Grid Scheduler for Independent Tasks", *International Journal of Computer Science Issues*, Vol. 8, 2011, pp. 296-301.
18. H. Izakian, B. T. Ladani, A. Abraham, V. Snasel, "A Discrete Particle Swarm Optimization Approach For Grid Job Scheduling", *International Journal of Innovative Computing, Information and Control*, Vol. 6, No. 9, 2010, pp. 1-15.
19. A. Revar, M. Andhariya, D. Sutariya, "Load Balancing in Grid Environment using Machine Learning-Innovative Approach", *International Journal of Computer Applications*, Vol. 8, No. 10, 2010, pp. 31-34.
20. G. D. Parmar, S. K. Mitra, "Performance Analysis of Unsupervised Probabilistic", *IJCA Special Issue on Computer Aided Soft Computing Techniques for Imaging and Biomedical Applications CASCT*, 2010, pp. 93-98.
21. J. MacQueen, "Some methods for classification and analysis of multivariate observations", *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967, Vol. 1, pp. 281-297.
22. K. V. Mardia, J. T. Kent, "Multivariate Analysis", *Academic Press*, 1980.
23. I. Breskovic, M. Maurer, V. C. Emeakaroha, I. Brandic, J. Brandic, "Towards Autonomic Market Management in Cloud Computing Infrastructures", *In Proceedings of CLOSER*, 2011.
24. J. A. Hartigan, *Clustering Algorithms*, John Wiley & Sons Inc, 1975.
25. S. Sauravjyoti, D. K. Bhattacharyya, "An Effective Technique for Clustering Incremental Gene Expression data", *International Journal of Computer Science Issues*, Vol. 7, No. 3, 2010, pp. 31-41.
26. G. Murugesan, Dr. C. Chellappan, "An Economic-based resource Management and Scheduling for Grid Computing Applications", *International Journal of Computer Science Issues*, Vol. 7, No. 2, 2010, pp. 20-25.
27. U. Karthick Kumar, "A Dynamic Load Balancing Algorithm in Computational Grid Using Fair Scheduling", *International Journal of Computer Science Issues*, Vol. 8, No. 1, 2011, pp. 123-129.
28. K. S. Chatrapati, J. U. Rekha, A.V. Babu, "Competitive Equilibrium Approach for Load Balancing a data Grid", *International Journal of Computer Science Issues*, Vol. 8, 2011, pp. 427-437.
29. S. R. Nimmagadda, P. Kanakamedla, V. B. Yaramala, "Implementation of Clustering Through Machine Learning Tool", *International Journal of Computer Science Issues*, Vol. 8, 2011, pp. 295-401.

30. T. R. V. Anadharajan, Dr. M. A. Bhagyaveni, "Co-operative Scheduled Energy Aware Load-Balancing technique for an Efficient Computational Cloud", International Journal of Computer Science Issues, Vol. 8, 2011, pp. 571-576.

**Seiiedeh Yasaman Rashida** received his B.S. in computer engineering from Sari Azad University, Mazandaran, in 2006, the M.S. in computer engineering from Arak Azad University, Markazi, in 2010. She is a teacher at Shirgah University in Mazandaran. Her research interests include grid computing and Fuzzy logic. She has some papers in these fields.

**Amir Masoud Rahmani** received his B.S. in computer engineering from Amir Kabir University, Tehran, in 1996, the M.S. in computer engineering from Sharif University of technology, Tehran, in 1998 and the PhD degree in computer engineering from IAU University, Tehran, in 2005. He is assistant professor in the Department of Computer and Mechatronics Engineering at the IAU University. He is the author/co-author of more than 80 publications in technical journals and conferences. He served on the program committees of several national and international conferences. His research interests are in the areas of distributed systems, ad hoc and sensor wireless networks, scheduling algorithms and evolutionary computing.

# Mobility Metrics Estimation and Categorization for SNET Protocols

Anita Sethi<sup>1</sup>, Prof. J. P. Saini<sup>2</sup>, Dr. Shailendra Mishra<sup>3</sup>,

<sup>1</sup> Uttarakhand Technical University,  
Dehradun, India

<sup>2</sup> M.M.M. Engineering College  
Gorakhpur, India

<sup>3</sup> Professor & Head, BCTKEC  
Dwarahat

## Abstract

In the performance assessment of an ad-hoc protocol, the protocol should be verified under genuine circumstances including, but not restricted to, an effective transmission range, inadequate buffer space for the storage of messages, illustrative data traffic models, and a mobility model. Deterioration of average speed as the simulation grows, a variance between the extensive duration distribution of nodes and the initial one, and sometimes the unpredictability of the model are the problems caused by simulation of mobility models. A mobility model emulates the actual world movement of mobile nodes and is essential component to simulation based analysis. So, satisfactory analytical and real demonstration of mobility is a very significant subject in simulation of mobile ad hoc networks. Our domino effects point out that the ad hoc protocols are certainly inclined by these mobility models. Extensive simulations have been conducted for different conditions of network density and node mobility for each of the four mobility models and also for different values of the degree of randomness parameter for the Gauss-Markov mobility model.

**Keywords:** *Random Waypoint Mobility Model, City Section mobility model, Manhattan mobility model, Gauss-Markov mobility model.*

## Introduction

Traces and syntactic are the two tactics for modeling of the mobility pattern. In trace-based models provide mobility patterns that are perceived in real-life systems and in it everything is deterministic whereas syntactic models represent the movements of mobile nodes accurately. Further Individual mobile movements and Group mobile movements are two categories of syntactic models. Random models based on statistical properties, models with temporal dependency influenced by their movement histories, models with spatial dependency, and models with geographical restrictions are four prime categories. The mobility pattern of the distinct mobile node is measured in case of Individual Mobility Models. In group mobility model, the supportive group movement of the mobile nodes performances in synchrony as a group, and reference random point group mobility model is a

model of this classification. The autoregressive mobility model contemplates mobility patterns of distinct nodes associating the mobility grades that may involve position, velocity, and acceleration at consecutive time instants. A synchronized movement job is performed by dynamic mobile nodes over (visually invisible) self-organized networks in nature in flocking and swarm mobility model. The virtual game-driven mobility model agreements with a distinct node or a cluster of mobile nodes based on user/player policies that are mapped from the real world to virtual agents interacting with each other or with groups of mobile users. In non-recurrent mobility model, the moving objects move in a totally unknown way without repeating the previous patterns, and these moving objects can be mobile nodes of the ad hoc network that constantly changes its topology or the continuously moving data arises in a broad variety of applications, including weather forecast, geographic information systems, air-traffic control, and telecommunications applications.

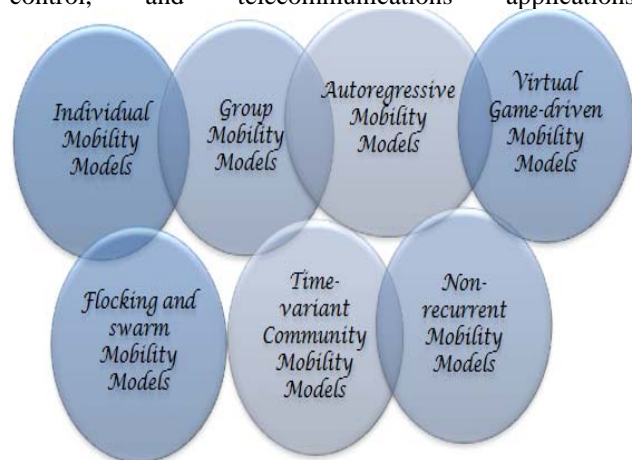


Figure 1 Classification of Mobility Models

The time-variant community mobility model captures the non-homogeneous behaviors in both space and time with inter-node dependency of the community of interest (COI)-based mobile network structure. The framework estimates the performance of spontaneous network routing protocols over diverse mobility patterns

that capture specific characteristics. The mobility models used in our analysis include the Random Waypoint, City Section mobility model, Gauss-Markov mobility model and Manhattan mobility model. Mobility metrics objective is to capture some of the aforesaid mobility characteristics and Connectivity graph metrics aim to study the effect of different mobility patterns on the connectivity graph of the mobile nodes. It has also been observed in previous works that under a given mobility pattern, routing protocols perform differently because each protocol differs in the basic mechanisms or “building blocks” it uses.

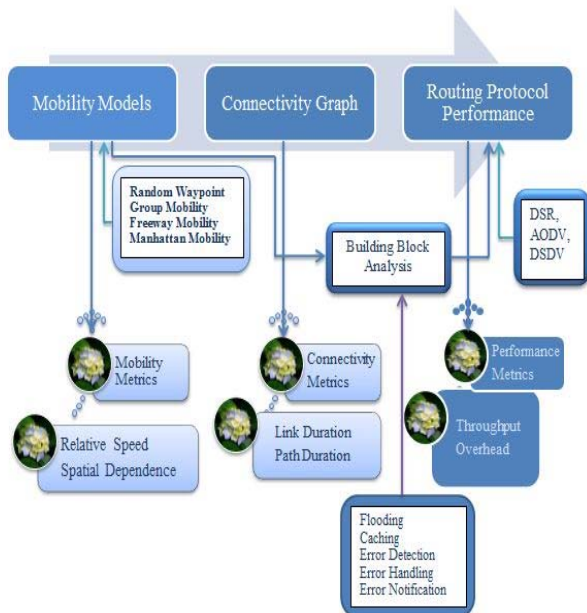


Figure 2 Framework of the Impact of Mobility on the Performance of Routing Protocols.

### Formulation of Mobility Models

The random characteristics of mobile nodes in a Spontaneous Network may comprise of a stochastic process, and every single node’s movement may involve a number of sequences of random length intervals called mobility epoch during which a node moves in a constant direction at a constant speed. The direction and speed of a mobile node may show a discrepancy in accordance to mobility benchmarks depending on the varieties of mobility models from epoch to epoch. In group mobility, the similar may be the case for a group of mobile nodes. Figure 3a illustrates the movement of a node over convinced epochs by an arbitrary node  $n$  from its position  $n$  to another position  $n'$  over an interval of length  $t$ . If we take responsibility that node  $n$  moves with a velocity  $V_{in}$  and direction  $\theta_{in}$  at epoch  $i$  and the duration of epoch  $i$  of node  $n$  is  $T_{in}$ , node  $n$  moves a distance of  $V_{in} T_{in}$  at an angle  $\theta_{in}$ . Let us define the distance traversed during time interval  $T_{in}$  by a mobile node at epoch  $i$  as an epoch mobility vector  $R_i^n = V_i^n T_i^n$ .

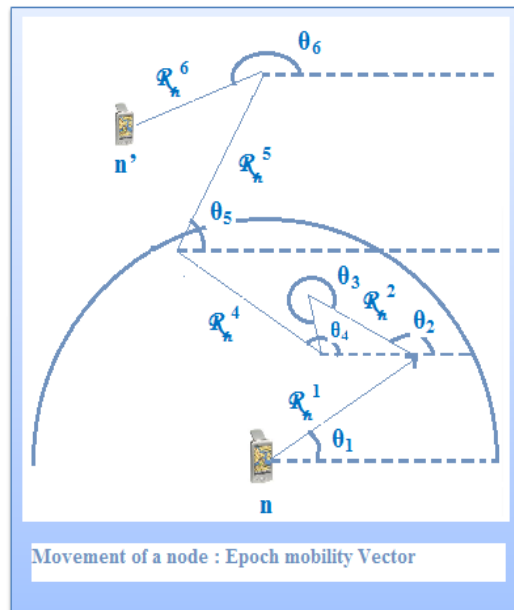


Figure 3a

In fact, Figure 3b shows the resulting epoch mobility vector  $R_i^n$  and it can be seen that this  $R_i^n$  is the vector sum of the individual epoch vectors. However, we need to examine the numerous parameters in order to articulate a mobility model.

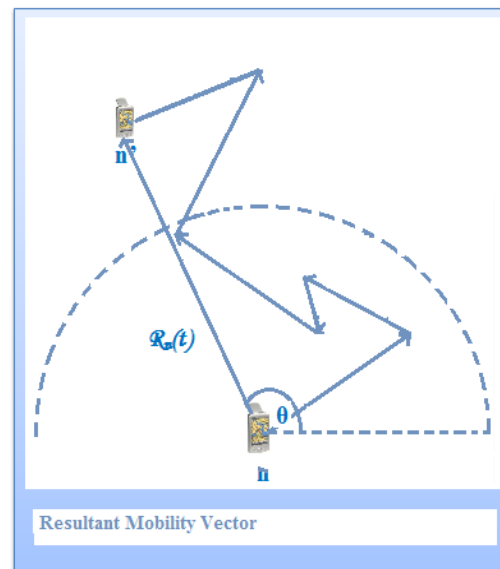


Figure 3b

Link availability is an appropriate metric in becoming died out links are the standards in spontaneous networks. Contrasting cellular systems where mobility is measured relative to a stationary base station, the mobility problem in ad hoc networks is more complex because both ends of each link are mobile. In a spontaneous network, every single communication path consists absolutely of wireless links, which frequently alteration independently of one another. The mobility of the nodes causes frequent link failures, triggering route recovery. Subsequently, Routing delay and the number of control packets are increased.

Mobility Matrices are used to represent the link and path stability, making it significant to investigate mobility metrics in random mobility models.

An energetic path with  $h$  hops between any two nodes at time  $t_0$ , the path availability  $\mathcal{A}(t, h)$  at time  $t$  is defined as the probability that the path exists at time  $t$ , given that it existed at time  $t_0$ .

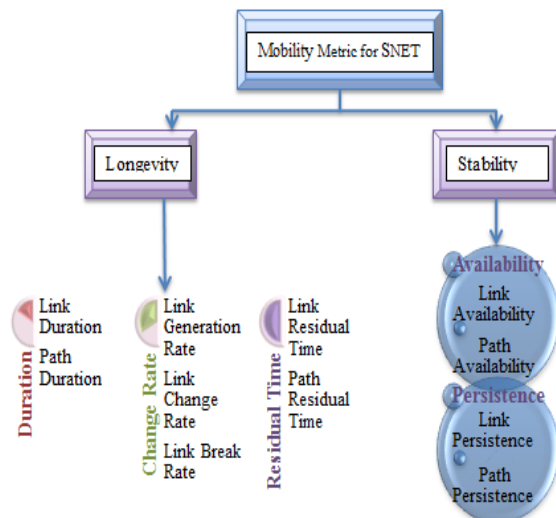


Figure 4 Mobility Metrics

An energetic path with  $h$  hops between any two nodes at time  $t_0$ , the path persistence  $\mathcal{P}(t, h)$ , as a function of time, is defined as the probability that the path will continuously last until at least time  $t$ , given that it existed at time  $t_0$ . An energetic path with  $h$  hops between any two nodes at time  $t_0$ , the path residual time,  $\mathcal{R}(h)$ , is the length of time for which the path will continue to exist until it is broken. The link residual time is denoted  $\mathcal{R} \triangleq \mathcal{R}(1)$ .

### Random Waypoint Mobility Model

Originally, the nodes are presumed to be positioned at random positions in the spontaneous network. The movement of every single node is self-governing of the other nodes in the network. The mobility of a specific node is defined as follows:

The node elects a random target position to travel with the velocity using which the node passages to this chosen position is uniform-randomly nominated from the interval  $[v_{min}, \dots, v_{max}]$ . The node travels in a straight line in a specific direction to the preferred position with the elected velocity. After accomplishment of the objective setting, the node may rest there for a definite time called the pause time. The node then carry on to choose another objective position and moves to that position with a new velocity chosen again from the interval  $[v_{min}, \dots, v_{max}]$ . The selection of each target location and a velocity to move to that location is independent of the current node location and the velocity with which the node reached that location.

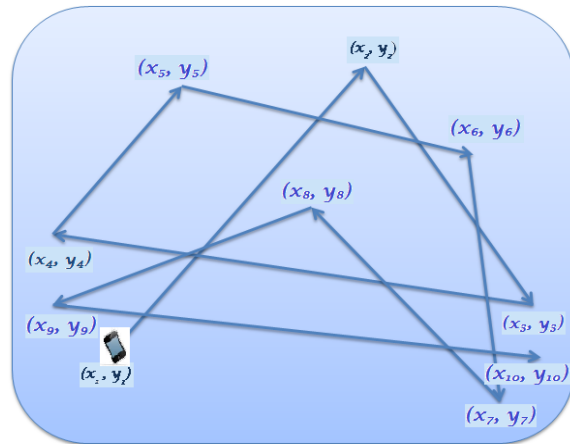


Figure 5 Random waypoint Mobility Model

### City Section Mobility Model

Primarily, the nodes are presumed to be arbitrarily located in the path intersections. Each one path is assumed to have a specific speed limit. Based on this speed perimeter and the block length, one can conclude the time it would take to move in the path. Every node positioned at a specific path intersection elects an arbitrary objective path intersection to move will experience the least amount of travel time. In case of two or more paths incur the least amount of travel time, the tie is broken arbitrarily. After attainment of the objective path intersection, the node may stay there for a pause time and then again choose an arbitrary objective path intersection to move. This procedure is repeated independently by each node.

### Manhattan Mobility Model

At first, the nodes are supposed to be haphazardly engaged in the path intersections. The movement of a node is decided one path at a time. Initially, every single node has equal probability of electing any of the paths leading from its primary position. After a node initiates to move in the elected direction and touches the successive path intersection, the subsequent path in which the node will move is chosen probabilistically. If a node can stay to move in the identical direction or can also change directions, then the node has 0.5 probability of staying in the identical direction, probability of 0.25 for fine-turning to the east/north and 0.25 probability of fine-turning to the west/south, depending on the direction of the prior movement. If a node has only two alternatives like situation when the node is in one of the four bounding paths of the network, then the node has an equal probability of discovering either of the dualistic options. If a node reaches any of four corners of network, then the node has no other choice except to explore that option..

### Gauss-Markov Mobility Model

In the beginning, the nodes are arbitrarily positioned and the movement of a node is self-determining in the network. Every node  $i$  is assigned a mean speed,  $\bar{S}_i$ ,

and mean direction  $\theta_i$  of movement. For every constant time period, a node calculates the speed and direction of movement based on the speed and direction during the previous time period, along with a certain degree of randomness incorporated in the calculation. The node is assumed to move with the calculated speed and in the calculated direction during every fixed time period. For a particular time instant,  $t_i^{a+1}$ , the speed and direction of a node  $i$  is calculated as follows:

$$S_i^{a+1} = \alpha * S_i^a + (1 - \alpha) * \overline{S}_i + \sqrt{1 - \alpha^2} * S^G(t_i^a)$$

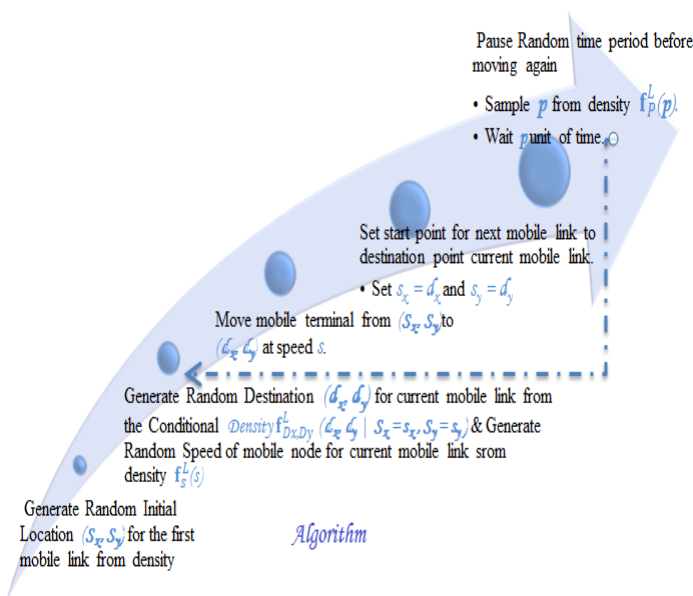
$$\Theta_i^{a+1} = \alpha * \Theta_i^a + (1 - \alpha) * \overline{\Theta}_i + \sqrt{1 - \alpha^2} * \Theta^G(t_i^a)$$

The parameter  $\alpha$  ( $0 \leq \alpha \leq 1$ ) is used to incorporate the degree of randomness while calculating the speed and direction of movement for a time period. The degree of randomness decreases as we increase the value of  $\alpha$  from 0 to 1. When  $\alpha$  is closer to 0, the degree of randomness is high, which may result in sharper turns. When  $\alpha$  is closer to 1, the speed and direction during the previous time period are given more importance (i.e., the model is more temporally dependent) and the node prefers to move in a speed and direction closer to what it has been using so far. Thus, the movement of a node gets more linear as the value of  $\alpha$  approaches unity. The terms  $S^G(t_i^a)$  and  $\Theta^G(t_i^a)$  are random variables chosen independently by each node from a Gaussian distribution with mean 0 and standard deviation 1. If  $(x_i^a, y_i^a)$  are the coordinates of node  $i$  at time instant  $t_i^a$ , the co-ordinates  $(x_i^{a+1}, y_i^{a+1})$  of the node at time instant  $t_i^{a+1}$  are given by:

$$X_i^{a+1} = X_i^a + [S_i^a * \cos(\Theta_i^a)]$$

$$Y_i^{a+1} = Y_i^a + [S_i^a * \sin(\Theta_i^a)]$$

### ALGORITHM



### Performance Metrics

The following performance metrics are evaluated:

**Percentage Network Connectivity:** Percentage network connectivity indicates the probability of finding an  $s-d$  path between any two nodes in the network for a given network density and level of node mobility. Measured over all the  $s-d$  sessions, this metric is the ratio of the number of static graphs in which there is an  $s-d$  path to the total number of static graphs in the mobile graph.

**Average Route Lifetime:** The average route lifetime is the average of the lifetime of all the static paths of an  $s-d$  session, averaged over all the  $s-d$  sessions.

**Average Hop Count:** The average hop count is the time averaged hop count of a mobile path for an  $s-d$  session, averaged over all the  $s-d$  sessions. The time averaged hop count for an  $s-d$  session is measured as the sum of the products of the number of hops for the static  $s-d$  paths and the corresponding lifetime of the static  $s-d$  paths divided by the number of static graphs in which there existed a static  $s-d$  path. For example, if a mobile path comprises of a **2-hop** static path  $p1$ , a **3-hop** static path  $p2$ , and a **2-hop** static path  $p3$ , existing in static graphs **1-3**, **4-7** and **8-10** respectively, then the time-averaged hop count of the mobile path would be  $(2*3 + 3*4 + 2*3)/10 = 2.4$ .

**Network Connectivity:** In case of minor-density networks, the inferior network connectivity achieved with the dual mobility models can be credited to the constrained motion of the nodes. In the instance of the Manhattan mobility model, the probabilistic behavior of direction selection after reaching each boarder is also a cause behind the lowest network connectivity detected for this mobility model among all the four mobility models. The amount of nodes dispersed in the area of the network may not be sufficient enough to connect any pair of source-destination nodes all the time. The direction of movement of the nodes is constrained close to the originally initialized mean direction of movement randomly chosen from  $[0...2\pi]$ , in the case of the Gauss-Markov mobility model. When there are rare nodes in the network, the limited movement of the nodes close to the mean direction of movement is a restraining factor for network connectivity. Together the Random way mobility model and the Gauss-Markov mobility model exhibit a significant growth in network connectivity, for all levels of node mobility once we raise the number of nodes in the network. This demonstrates the fact that the randomness related with the mobility models guarantees that any combination of nodes will persist connected, provided we have at least a sensibly superior number of nodes, irrespective of the different levels of node mobility.

### Simulation Results

#### Minimum hop mobile path

We now discuss the time averaged hop count per minimum hop path (refer Figure 8) and the average route lifetime (refer Figure 9) of the minimum hop paths

determined as the constituent paths of the Minimum Hop Mobile Path under the four mobility models.

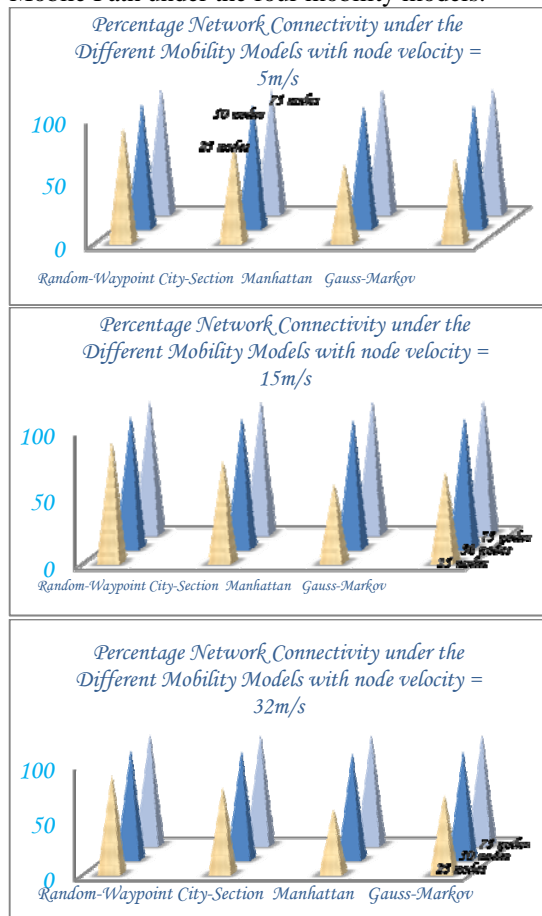


Figure 7 Percentage Network Connectivity

### Average hop count per minimum hop path

The average hop count per minimum hop path determined for the two SNET mobility models and the Gauss-Markov model is considerably larger than the hop count per minimum hop path determined for the Random Waypoint mobility model. The relatively larger hop count can be attributed to the constrained mobility of the nodes under these three mobility models. The minimum hop paths in the street networks are most likely not to exist on or close to the straight line between source and destination nodes. Similarly, due to the temporal dependency associated with the Gauss-Markov mobility model, one cannot always find minimum hop paths lying on a straight line connecting the source and destination nodes. Based on our observations in Figures 8 we can arrive at the following ranking of the four mobility models in the increasing order of the magnitude of the hop count for the minimum hop paths: Random Waypoint model, City Section model, Gauss-Markov model and the Manhattan model.

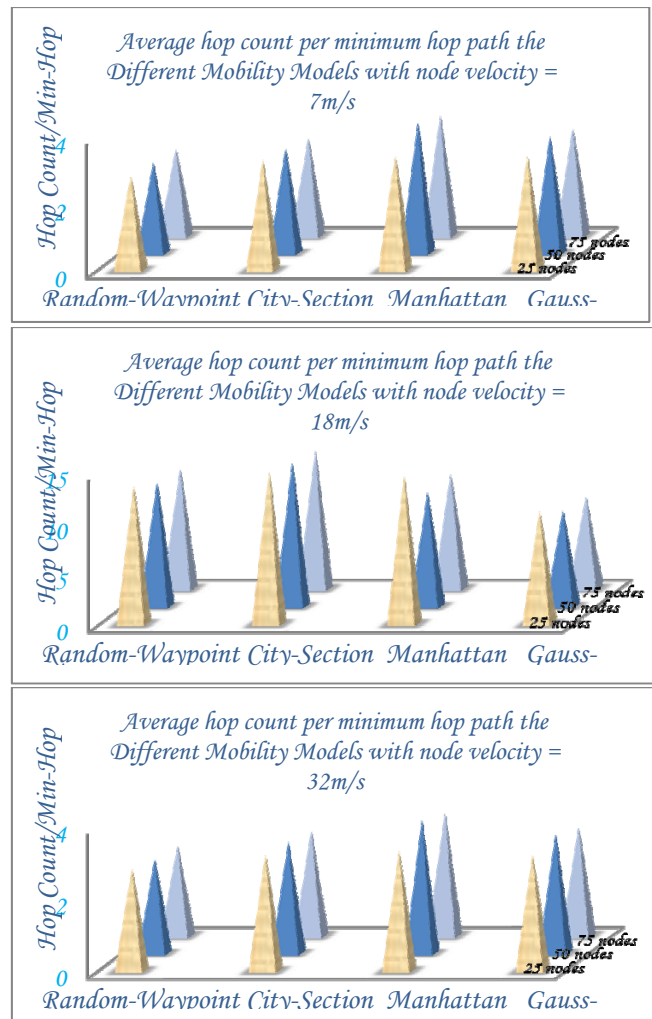


Figure 8 Average Hop count per min. hop Path

The average hop count per minimum hop path with a particular node velocity, under the Manhattan mobility model, Gauss-Markov mobility model and the City Section mobility model is respectively about 0.19, 0.17 and 0.14 more than that experienced for the Random Waypoint mobility model in minor-density networks. In moderate and high-density networks, the average hop count per minimum hop path under the City Section, Gauss-Markov and Manhattan mobility models is respectively about 18%, 25% and 40% more than that incurred for the Random Waypoint mobility model. We also observe that with increase in network density, the average hop count per minimum hop path for the Random Waypoint mobility model, City Section mobility model and the Gauss-Markov mobility model decreases (by a factor of 5%-10%).



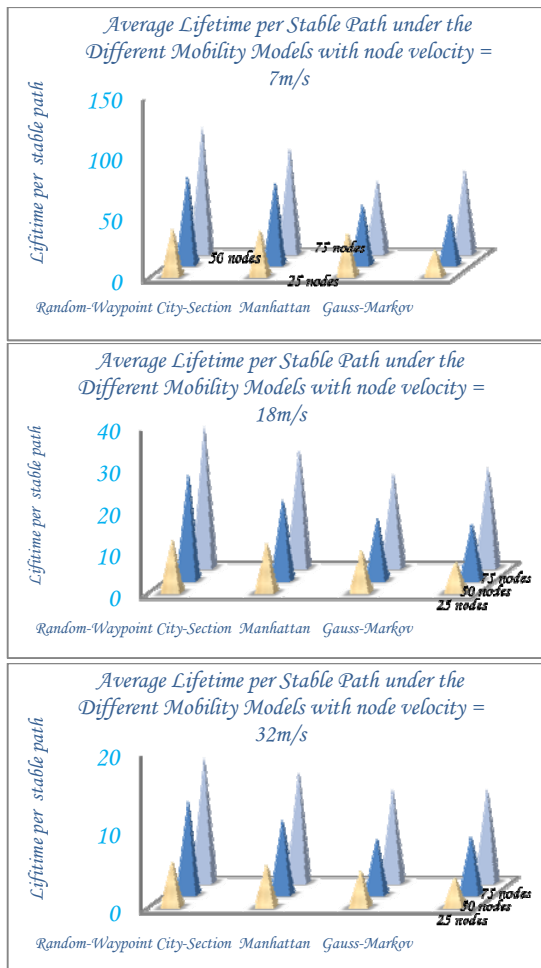


Figure 9 Average Lifetime per Stable Path

On the other hand, with increase in network density, the average hop count per minimum hop path for the Manhattan mobility model remained the same or even sometime increases up to 14%. This can be attributed to the significant increase in the network connectivity for the Manhattan mobility model with increase in network density, but at the cost of increase in hop count. Given the particular self-conscious constraints and random behavior of node movement under the Manhattan mobility model, with respect to source and destination, supplementary intermediary nodes have to be accommodated in the s-d paths.

### Conclusions

Like the model of quality of service is carry out by a set of constraints; the mobility model will be carry out by a set of metrics. The study signposts that the significant constraints for the precision of mobility metrics are the node density dissemination and the stability. The techniques of mobility metric calculation have straight emotional impact on the metric accuracy. The more genuine is a mobility model, the superior is the number of hops in the minimum hop routes and smaller is the lifetime of stable routes determined under the mobility model. The Random Waypoint model yielded the lowest hop count for

minimum-hop routes and the largest lifetime for stable routes. On the other hand, more realistic mobility models such as the Gauss-Markov model and the Manhattan model yield a relatively larger number of hops for minimum-hop routes and a relatively smaller lifetime for stable routes. For a specified scenario of network density and node mobility, the average hop count of a Minimum Hop Mobile Path is less significant than the average hop count of a Stable Mobile Path; the average route lifetime of a Stable Mobile Path is more than the average route lifetime of a Minimum Hop Mobile Path.

### References

- [1] S. A. Kulkarni and G. R. Rao, "Mobility Model Perspectives for Scalability and Routing Protocol Performances in Wireless Ad hoc Networks," Proceedings of the First International Conference on Emerging Trends in Engineering and Technology, pp. 176 – 181, July 2008.
- [2] T. Camp, J. Boleng and V. Davies, "A Survey of Mobility Models for Ad Hoc Network Research," Wireless Communication and Mobile Computing, Vol. 2, No. 5, pp. 483-502, September 2002.
- [3] F. Bai, N. Sadagopan and A. Helmy, "IMPORTANT: A Framework to Systematically Analyze the Impact of Mobility on Performance of Routing Protocols for Ad hoc Networks," Proceedings of the IEEE International Conference on Computer Communications, pp. 825-835, March-April, 2003.
- [4] J. Ariyakhajorn, P. Wannawilai and C. Sathitwiriawong, "A Comparative Study on Random Waypoint and Gauss-Markov Mobility Models in the Performance Evaluation of MANET," Proceedings of the International Symposium on Communications and Information Technologies, pp. 894-899, September 2006.
- [5] M. Abolhasan, T. Wysocki and E. Dutkiewicz, "A Review of Routing Protocols for Mobile Ad hoc Networks," Elsevier Ad Hoc Networks, Vol. 2, No. 1, pp. 1-22, January 2004.
- [6] S. P. Chaudhri, J. Y. L. Boudec, M. Vojnovic, "Perfect Simulations for Random Trip Mobility Models," Proceedings of the 38th Annual Symposium on Simulation, pp. 72-79, San Diego, USA, April 2005.
- [7] M. I. M. Saad and Z. A. Zukarnain, "Performance Analysis of Random-based Mobility Models in MANET Routing Protocol," European Journal of Scientific Research, Vol. 32, No. 4, pp. 444-454, 2009.
- [8] S. A. Kulkarni and G. R. Rao, "Mobility Model Perspectives for Scalability and Routing Protocol Performances in Wireless Ad hoc Networks," Proceedings of the First International Conference on Emerging Trends in Engineering and Technology, pp. 176 - 181, July 2008.
- [9] B. Divecha, A. Abraham, C. Grosan and S. Sanyal, "Impact of Node Mobility on MANET Routing Protocols Models," Journal of Digital Information Management, Vol. 4, No. 1, pp. 19 - 23, February 2007.
- [10] G. Ravikiran and S. Singh, "Impact of Mobility Models on the Performance of Routing Protocols in Ad hoc Wireless Networks," Proceedings of the 59th IEEE Vehicular Technology Conference (VTC 2004-Spring), Vol. 4, pp. 2185 - 2189, May 2004.
- [11] F. Bai, N. Sadagopan and A. Helmy, "The IMPORTANT Framework for Analyzing the Impact of Mobility on Performance of Routing Protocols for Ad hoc Networks," Elsevier

Ad hoc Networks, Vol. 1, No. 4, pp. 383 - 403, November  
2003.

# Using Petri Nets For Resource Management Modeling In The Operating Systems

Adalat Karimov<sup>1</sup>, Shahram Moharrami<sup>2</sup>

<sup>1</sup>Department of Information Economy and Technologies,  
Azerbaijan State Economic University,

<sup>2</sup>Department of Computer Engineering, Parsabad Moghan Branch,  
Islamic Azad University, Iran,

## Abstract

Nowadays, with advances in computer science and increase in processor speed, modeling methods have found extensive applications in industrial fields. Petri Nets are one of these modeling methods. Petri Nets are based on graph theory and are applied specifically for concurrent and asynchronous applications. As executable models, they are capable of graphical description of complicated systems. On the other hand, development of hardware and other peripheral computer resources and development of various computer software systems call for efficient and powerful operating systems, so that users can use the software and hardware items in an effective manner. The purpose of this article is to study the application of Petri Nets for modeling resource management in operating systems with the aim of optimal utilization of resources and Deadlock Avoidance in the Operating Systems.

**Keywords** Petri Nets ,Resources ,Deadlock Avoidance, Operating System, Place, Transition.

## 1- Introduction

A model is a simple and understandable representation of the structure and behavior of the system under study which in most cases can be expressed by mathematical formulas. Decision rules can be obtained directly from a pattern [10,11]. In fact, using the model, we can acquire knowledge and information about the modeled phenomenon without experiencing the costs and risks associated with the real phenomenon. In the same way, Petri Nets are used for acquiring information about structure and function of modeled systems. Petri Nets were developed by Mr. Carl Adam Petri in 1962 [1]. His research focused on information systems. Numerous groups were formed in Germany and several other countries to conduct research projects on applications of Petri Nets. In this article, after a brief introduction to Petri Nets, we will investigate the problem of resource

management modeling in the operating system using Petri Nets.

## 2- Petri Nets

Petri Nets are based on graph theory. They are executable and have the capability for graphical description of complicated systems. The theory of Petri Nets has grown in two directions [4]:

- a. Applied: the applied theory of Petri Nets;
- b. Theoretical: the pure theory of Petri Nets.

The objective of the developments in the applied domain is to provide essential tools, techniques, and relationships for application of Petri Nets. In fact, strong theory is a prerequisite for more effective applications. Developments in the theoretical direction consider that Petri Nets are useful for real world problems. Accordingly, we take into account both theoretical and applied aspects in this article. First, we define some formulas and then we describe the power and applications of this tool.

### 2.1 Basic Definition

In this section, we present the official definition of hierarchical Petri Nets. This definition is required for a basic understanding and study of Petri Nets. Our formulation is based on the theory of sets. In fact, these nets provide a new class of machines called Petri Net automata.

### 2.2 Structure Of Petri Nets

Since Petri Nets (PN) are a special form of graphs, we start with a few basic concepts from the graph theory [5].

**Definition 1.** A graph consists of two components, the nodes and the edges, and a method for connecting these elements[6].

In other words, a graph  $G(V, E, Q)$  consists of a non-empty set  $V$ , which is called the *set of nodes* of the graph, a set  $E$  called *edges* of the graph, and a mapping  $Q$  from the set  $E$  onto the set of pairs  $v$ . If the pair of nodes that are connected by an edge is an ordered pair, then the edge is directed and an arrow is placed on the edge to denote this fact. If all edges of a graph are directed, the graph is called a directed graph (a digraph). Two nodes that are connected by an edge of the graph are called *adjacent nodes*. Nodes need not necessarily be denoted by points; they can be shown using circles, bars, boxes, or any other conventional shape. When a graph contains directed parallel edges (edges that connect an identical pair of nodes), that graph is called a multi-graph. A few instances of multi-graphs are shown in Fig1. The property of Petri Nets as graphs is that they are bi-partite graphs, i.e., they have two kinds of nodes. To differentiate these two kinds of nodes, different shapes are used to represent them. The first type of nodes is called *place nodes* and they are shown by a circle or ellipse. The second type of nodes are called *transition nodes* and are denoted by a solid bar or rectangle. The edges of Petri Nets are called *arcs* and are always directed. A bi-partite graph has special properties. An edge can only connect two nodes of different types. So the edges can be arcs from a place to a transition or from a transition to a place. It is not possible for an arc to connect a place to another place or a transition to another transition[9]. Fig1 shows two instances sample graphs.



Fig 1: Two sample graphs.

**Definition 2.** A Petri Net is composed of four components [1,14]:

- $P = \{p_1, p_2, \dots, p_k\}, k > 0$  is a finite set of places,
- $T = \{t_1, t_2, \dots, t_l\}, l > 0$  is a finite set of transitions (with  $P \cup T \neq \emptyset$  and  $P \cap T = \emptyset$ ),
- $I: P \times T \rightarrow N$  is an input function that specifies weights of arcs directed from places to transitions,
- $O: P \times T \rightarrow N$  is an output function that specifies weights of arcs directed from transitions to places.

$I/O$  functions are the connecting bridge of transitions ( $T$ ) and places ( $P$ ), and connect  $T, P$  pairs to each other. The input function determines the set of input

places  $I(t_i)$  for each transition  $t_i$ , while the output function determines the set of output places  $O(t_i)$  for each transition  $t_i$  [7]. The structure of a Petri Net is defined with places, transitions, and input and output functions. Fig2 shows two instances of Petri Nets.

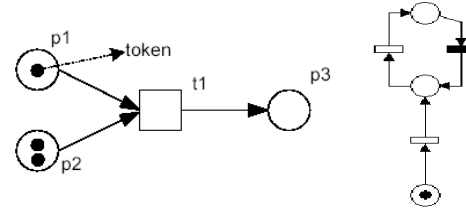


Fig 2: Two sample Petri Nets.

### 2.3 Petri Net Notations

Notation  $m$  marks the placement of a series of tokens on the places of the Petri Net. Each token has an initial concept of  $T, P$  in the Petri Net. After the tokens are defined and put in places ( $P$ ) of Petri Net, the number and state of tokens can change during execution of a Petri Net. In other words, tokens are used to define the execution of the Petri Net.  $M = (C, m)$  is a Petri Net, where  $C = (P, T, I, O)$  and  $m$  is a notation. This is often written as [6]:

$$M = (P, T, I, O, m)$$

Vector  $m$  shows the number of tokens for each place  $P_i$ .  $m_i$  is the number of token at place  $P_i$ , i.e.  $M(P_i) = m_i$ .

### 2.4 Execution Rules Of Petri Nets

Petri Nets are executed by firing transitions[8]. A net is executed by means of separate indexed tokens. Tokens are put in places and control the execution of net transitions. A transition  $T$  is executed by removing tokens from input places and creating new separate tokens in output places. A transition fires only if it is permissible or active, and it is permissible if each of input places has at least the number of tokens equal to input arcs of that transition (which are arrows from place  $P$  to transition  $T$ ). The number of tokens needed for input arcs to activate a transition  $T$  are called activating tokens. Transition  $T_i$  in a signed Petri Net  $C = (P, T, I, O)$  with sign  $m$  is activated when the number of provided tokens  $p_i$  is greater than or equal to the minimum number of incoming arcs to  $T_i$  [1,12,13].

**Example.** In Fig3,  $t_1$  is fired when there are at least one token in  $p_1$  and  $t_3$  is fired when there are one token in  $p_3$ . When a transition is fired, it pushes out all activator tokens and sends them to output places.

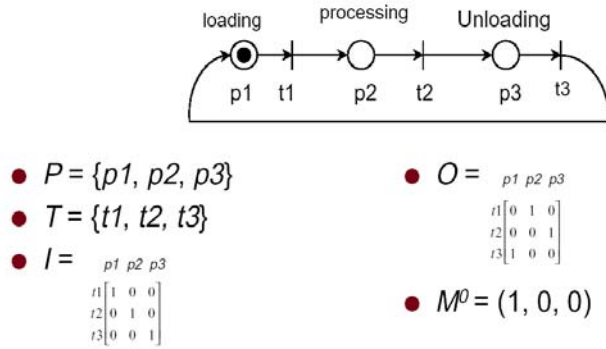


Fig 3: Places, transitions, and transition conditions.

### 3. Operating System, Rules And Methods Of Modeling

The operating system is the main program that acts as an interface between the user and computer hardware, and is usually the first program that is loaded into the memory after booting up the computer [2]. The main part of the operating system that performs its important tasks is called the *kernel*. The kernel is always running on the computer. The operating system has two main duties:

- (a) It simplifies working with the computer. In other words, the user is not forced to deal with hardware issues or to call hardware system operations like writing to the disk directly.
- (b) The second task of the operating system is resource management. Resource management enables optimal and cost-effective utilization of system resources (processors, memories, disks, etc.).

The operating system works as a resource manager and allocates resources to programs according to their requirements. Evidently, the number of requested resources should not be more than the total number of available resources. If each process in a set waits for an event that can occur only by another process in that set, a deadlock state ensues [2].

Our goal in modeling resource management is to detect and repair deadlock conditions in the system. Therefore, we must model requirements in a way that shows their role in the occurrence of deadlock. In fact, the effects and results of requests on the system are very important. On the other hand, modeling requests is not separate from modeling available resources. Therefore, the role of requirements in available resources should be elucidated clearly. If a program abuses a requirement, it can accentuate its role in creating deadlock. As a matter of fact, the requirements lead to enhancement of available resources. Therefore, in the modeling, the available resources should be indicated in addition to requirements [3].

### 4. The Proposed Model

The proposed model consists of two essential parts. The first part includes characteristics of the requirements for the processes and the second part includes the characteristics of available resources. In this modeling, a separate Petri net is used for each process, consisting of a transition and two places. One place includes the tokens denoting the characteristics of resource requirements of the process (including their number and type). The other place includes tokens denoting characteristics of available resources (their type and number). Fig4 shows a general schematic representation of a basic Petri net.

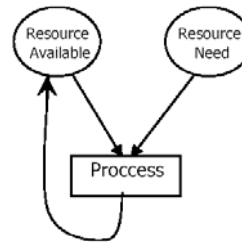


Fig 4: General sketch of a basic Petri net.

However, each basic Petri net related to a process has a transition in addition to the two places. The inputs to this transition come from the above-mentioned places and consist of resource requirements of the process and the available resources. The outputs are related to available resources, where allocated resources are added to available resources after termination of the process. Each transition has a conditional expression for activation, so that the transition takes effect if the condition is met. The operators in these conditional expressions consist of the usual logical operators, and their operands are the number of input tokens to the transition. In fact, the transition shows the possibility of the execution of the process. The transition reviews the possibility of the process according to available resources, and if the process is possible, the transition is activated and the process is executed. The allocated resources return to available resources after termination of the process. This transition prevents the occurrence of a deadlock in the system.

### 5. Modeling Deadlock Concepts

An important concept associated with resource allocation is deadlock. Within Petri Net theory an important question is “does the net deadlock?” A Petri Net is deadlocked iff no transitions are enabled. The first subsection presents two simple net models that provide a means to introduce the concept of deadlock. In the second subsection a graphical model of the bankers algorithm and deadlock detection algorithm is discussed. This model is a completely different model from the first

two and exhibits a very different use of a Petri Net model for the same general subject area.

### 5.1. Introducing Deadlocks With Petri Net

The following model is used to introduce the concept of deadlock. Shown in Fig5 is a simple Petri Net model of a resource allocation scheme with only two instances of a single kind and two processes that require both resources before completing. The place R represents the single resource and each token represents an instance of the resource. The remaining places represent the thread of control of the two processes. The transitions  $t_1$  and  $t_2$  respectively represent the events of process a and process b requesting an instance of the resource. The transitions  $t_3$  and  $t_4$  represent a second request from each. Finally, transitions  $t_5$  and  $t_6$  model the releasing of both resource instances. This net can deadlock if the following firing sequence occurs:  $t_1, t_2$ . At this point no transitions are enabled. Although resource allocation graphs are useful, the Petri Net model adds the dynamic aspects explicitly, i.e. the possible sequences of events leading up to a deadlock. One may begin to examine what firing sequences cause deadlocks.

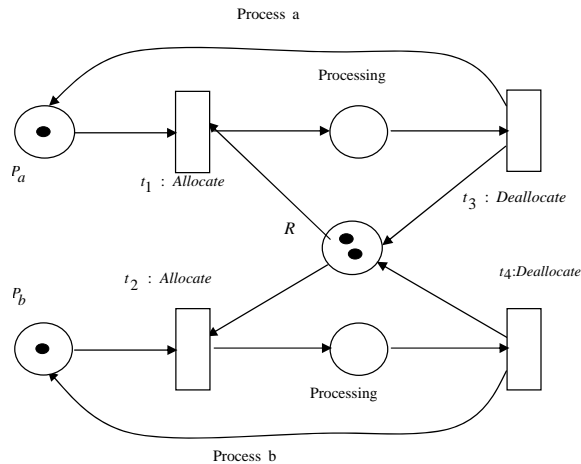


Fig 6: Resource allocation net model 2.

### 5.2. A Model For Deadlock Avoidance And Detection Algorithms

Deadlock avoidance and detection are two other major subtopics. Typically, the banker's algorithm is presented when discussing deadlock avoidance.

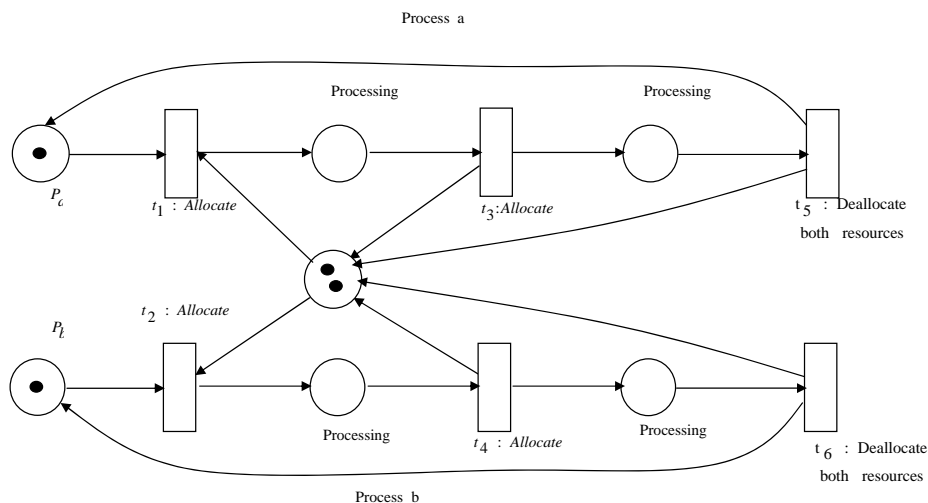


Fig5. Resource allocation petri net model 1.

Once the first model is understood and the concept of deadlock is introduced, the subject of deadlock prevention and the related costs can be discussed. To illustrate the prevention protocol that calls for processes to allocate all their resources before execution begins one could use the following Petri Net in Fig6. One can see that if  $t_1$  fires, then  $t_2$  and  $t_4$  cannot be enabled until  $t_3$  has been fired. This models the fact that process  $a$  has all the resources and process  $b$  cannot make any progress. Following the firing of  $t_3$ , both  $t_1$  and  $t_2$  are enabled.

The banker's (safety) algorithm checks for the safeness of allocating resources to a process  $p_i$  each time a request is made by process  $p_i$ . A similar algorithm is used for detecting deadlocks. The deadlock detection algorithm is run periodically. The output of the detection algorithm is either a sequence of processes, which indicates how the system of processes may complete, and thus indicate that the system is not deadlocked, or the indication that the system is deadlocked.

A Petri Net model for the deadlock detection algorithm is presented next for a system of processes with several instances of each resource type. In the following

discussion we use  $X$  as an unconventional index variable to emphasize that resources are indexed by letters rather than non-negative integers. Given  $n$  processes and  $m$  resource types, the algorithm uses an  $n \times m$  allocation matrix, *Allocation*, whose entry,  $Allocation[i, X] = k$ , indicates that  $k$  instances of resource  $X$  are allocated to process  $i$ . Similarly, an  $n \times m$  request matrix, *Request*, has entries,  $Request[i, X] = k$ , that indicate process  $i$  is currently requesting  $k$  instances of resource  $X$ . An  $m$ -vector, *Available*, whose entry,  $Available[X] = k$ , indicates  $k$  instances of resource  $X$  are currently available. The generic net model is described as follows. There is a process place and transition pair for each of the processes in the system (each correspondingly subscripted). There is also a resource place for each resource type. Let  $p_i$  and  $p_x$  denote process and resource places, respectively. The arc set  $F$  and labeling function  $W$  consist of:

arcs  $(P_i, t_i)$ , each with weight one, for each process  $i$ ;

arcs from transitions  $t_i$  to resource places  $p_x$  with weight label  $k$  iff  $Allocation[i, X] = k$  and  $k > 0$ ;

and arcs from resource places  $p_x$  to transitions  $t_i$  with weight label  $k$  iff  $Request[i, X] = k$  and  $k > 0$ .

In the algorithm, a Boolean  $n$ -vector, *Finish*, is used to keep track of what processes were examined and can have their requests met.  $Finish[i] = true$  when the  $i^{th}$  process has been examined by the algorithm and its request can be met. Initially,  $Finish[i] = false$  for all  $i$  where  $1 \leq i \leq n$ .

The place markings correspond to *Finish*. A process place  $P_i$  is marked with one token iff  $Finish[i] = false$ , and marked with zero tokens otherwise. After a transition  $t_i$  fires, a token is removed from process place  $P_i$ . This corresponds to setting  $Finish[i]$  to true. The meaning of firing a transition is discussed shortly. The resource places  $p_x$  are marked with  $k$  tokens iff  $Available[X] = k$ . In addition to the data structures discussed above, the algorithm also uses an  $m$ -vector called *Work*, which is modified during the execution of the algorithm. *Work* is initialized by the *Available* vector. At each point in the execution of the algorithm, the value  $Work[X]$  corresponds to the number of tokens in  $p_x$ . When  $Request[i, X] \leq Work[X]$  for each resource kind  $X$  and  $Finish[i] = False$ , the transition  $t_i$  is correspondingly enabled. This corresponds to the situation where the requests of the  $i^{th}$  process can be met. Note that more than one request might be met. Thus, several transitions might be enabled. One is nondeterministically chosen. If process  $i$  is chosen, then the *Work* vector is updated through vector addition of

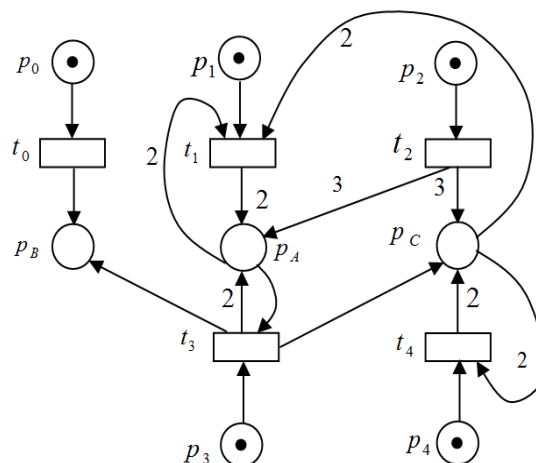
itself with the  $i^{th}$  Allocation row. This corresponds to the firing of transition  $t_i$ .

Suppose that five processes,  $\{p_0, p_1, \dots, p_4\}$ , and three resource types,  $\{A, B, C\}$ , are given. Resources  $A, B$ , and  $C$  have 7, 2, and 6 instances, respectively. Request matrix is *Request*. Suppose that the matrix representing the state of resource allocation at same time  $T$  are shows in table 1.

**Table 1:** Processes, Resource types and Requests

	Allocation				Request				Available		
	A	B	C		A	B	C		A	B	C
$P_0$	0	1	0		0	0	0	0	0	0	
$P_1$	2	0	0		2	0	2				
$P_2$	3	0	3		0	0	0				
$P_3$	2	1	1		1	0	0				
$P_4$	0	0	2		0	0	2				

The following net model in Fig7 allows one to graphically depict the deadlock detection algorithm. One begins to arbitrarily (nondeterministically) fire enabled transitions until no transitions are enabled.



**Fig7:** Deadlock Detection Model

Note that in this model, this does not mean the modelled system is deadlocked. Instead, the deadlock in the modeled system is detected if a token remains in any of the initially marked process places (as opposed to the resource places) and no transition is enabled. If all these process places are unmarked, then that firing sequence corresponds to a corresponding process completion sequence. For the above example  $t_0, t_2, t_3, t_1, t_4$  is one possible firing sequence. This indicates that the above system is not in a deadlocked state. Another example of the many possible firing sequences is:  $t_2, t_4, t_0, t_3, t_1$ .

The Petri Net model presented in this subsection is an alternative graphical technique for detecting deadlocks, and, as shown in this paper, can be used to model many other concepts in operating systems.

## 6. Conclusion

Petri Nets are an appropriate tool for modeling complicated systems, and are very useful for studying concurrency and uncertainty. Nowadays, modeling and simulation are used extensively in industrial fields. In fact, an industrial system can be studied before its creation and this is very cost-effective from the economic and time point of view. Petri Nets enable us to study the various components of the systems near each other. Using Petri Nets in the operating system research, we can investigate stability and concurrency and prevent the occurrence of deadlock. Using this method, which is a new method for modeling resource management, we can model complicated processes that require various resources.

## 7. References

- [1] Shams, Fereydun, Zinati Safa, An *algorithm* for modeling expert systems using fuzzy colored Petri Nets, M.Sc. degree thesis, Shahid Beheshti University, July 2008.
- [2] Silberschatz, Abraham, Operating System Concepts, translated by Atamazhuri, Parisima, Tehran, Iran: Ashian 1999.
- [3] Jalili, Rasul, Modeling vulnerabilities of computer networks using colored Petri Nets, Eleventh International Computer Conference of Iranian Computer Association, Computer Science Research Institute, Tehran, Iran, 2005.
- [4] Barzamini, Ruhollah, Petri Nets and modeling internet-based remote control systems using Petri Nets, The First National Conference on Computer Engineering, Islamic Azad University, Lahijan, Iran, 2007.
- [5] James L. Peterson, "Petri Nets", computing surveys, Vol 9, No.3, September, 1977.
- [6] J. Peterson, Petri Net theory and the Modeling of Systems, prentice-Hall, N.J., 1981.
- [7] Kurt Jensen, "A Brief Introduction to Coloured Petri Nets", Computer science Department, University of Aarhus, Denmark. 1997.
- [8] T. Murata, Petri Nets: "Properties, Analysis And Applications", Proceedings of IEEE 77 (1989) 540–541.
- [9] Burcin Bostan-Korpeoglu, Adnan Yazici, A Fuzzy Petri Net Model For Intelligent Database, Data & Knowledge Engineering (2006), Elsevier, 2006.
- [10] A. Karimov, S. Moharrami, "Automatic Classification with Neural Networks Using New Decision Rule", International Journal of Applied Mathematics & Statistics, Vol. 19, No. D10, 2010, pp. 90-96.
- [11] A. Karimov, S. Moharrami, "On Approaches To Automatic Classification Of Object States Based On New Decision Rule", in First International Conference on Soft Computing Technologies in Economy, November 2007, Baku, Azerbaijan, Vol. 1, pp. 107-116.
- [12] K. Jensen, "Colored Petri nets (CPN)", <http://www.daimi.au.dk>.
- [13] Burcin Bostan-Korpeoglu, Adnan Yazici, A Fuzzy Petri Net Model For Intelligent Database, Data & Knowledge Engineering (2006), Elsevier, 2006.
- [14] Motameni, H., et al. "Using Markov Theory For Deriving Non-Functional Parameters On Transformed Petri Net From Activity Diagram", proc of software engineering conference (russia), 16-17 November 2006, Moscow, Russia, (presented).



# MPLS - A Choice of Signaling Protocol

Muhammad Asif<sup>1</sup>, Zahid Farid<sup>2</sup>, Muhammad Lal<sup>3</sup>, Junaid Qayyum<sup>4</sup>

<sup>1</sup>Department of Information Technology and Media (ITM), Mid Sweden University  
Sundsvall 85170, Sweden.

<sup>2</sup>School of Electrical and Electronics Engineering, Universiti Sains Malaysia (USM)  
Penang 14300, Malaysia.

<sup>3,4</sup>Department of Information Technology, Gandhara University of Sciences  
Peshawar 25000, Pakistan.

## Abstract

Multi Protocol Label Switching (MPLS) is a core networking technology that operates essentially in between Layers 2 and 3 of the OSI model; for this reason, MPLS has been referred to as operating at Layer 2.5. MPLS can overlay existing technologies such as ATM (Asynchronous Transfer Mode) or Frame Relay, or it can operate in an entirely IP native environment; this can allow users to take advantage of existing CPE (Customer Premises Equipment) while making a move towards converging all network traffic, such as data, video and voice, at a pace that users can accommodate and afford. MPLS provides its users a number of advantageous features such as traffic engineering, network convergence, failure protection, and the ability to guarantee Quality of Service (QoS) over IP. MPLS Vans take advantage of the inherent characteristics of MPLS to provide secure data networking, typically for business users, in conjunction with other VPN technologies to help increase scalability while keeping costs at a manageable level. This paper should help to provide a basic understanding of MPLS technology, its advantages and limitations, and its application as an IP VPN. This paper covers MPLS, Label Distribution, Explicit Routes, Constrained Routes, Resource Reservation, Traffic Engineering, Service Level Contracts, Virtual Private Networks and Modern Networks needs. Our Next papers will focus on MPLS Traffic Engineering Overview and Differences and Similarities between RSVP and CR-LDP.  
**Keywords** Multi Protocol Label Switching (MPLS), OSI model, ATM (Asynchronous Transfer Mode), CPE (Customer Premises Equipment), Quality of Service (QoS), VPN (Virtual Private Networks), Traffic Engineering, Resource ReSerVation Protocol (RSVP), Constraint-based Routed Label Distribution Protocol (CR-LDP).

## 1. Introduction

MPLS is a new technology that offers to open up the Internet by providing many additional services to applications using IP. MPLS forwards data using labels that are attached to each data packet. These labels must be distributed between the nodes that comprise the network. Many of the new services that ISPs want to offer rely on Traffic Engineering functions. [1] There are currently two label distribution protocols that provide support for Traffic Engineering: Resource ReSerVation Protocol (**RSVP**) and Constraint-based Routed Label Distribution Protocol (**CR-LDP**). Although the two protocols provide a similar level of service, the way they operate is different, and the detailed function they offer is also not consistent. Hardware vendors and network providers need clear information to help them decide which protocol to implement in a Traffic Engineered MPLS network. Each protocol has its champions and detractors, and the specifications are still under development. Recognizing that the choice of label distribution protocol is crucial for the success of device manufacturers and network providers, this White Paper explains the similarities and important differences between the two protocols, to help identify which protocol is the right one to use in a particular environment. Data Connection's DC-MPLS family of portable MPLS products offers solutions for both the RSVP and CR-LDP label distribution protocols. Multi-Protocol Label Switching (MPLS) is a new technology that will be used by many future core networks, including converged data and voice networks. [2] MPLS does not replace IP routing, but will work alongside existing and future routing technologies to provide very high-speed data forwarding between Label-Switched Routers (LSRs) together with reservation of bandwidth for traffic flows with differing

Quality of Service (QoS) requirements. MPLS enhances the services that can be provided by IP networks, offering scope for Traffic Engineering, guaranteed QoS and Virtual Private Networks (VPNs). The basic operation of an MPLS network is shown in Figure 1.

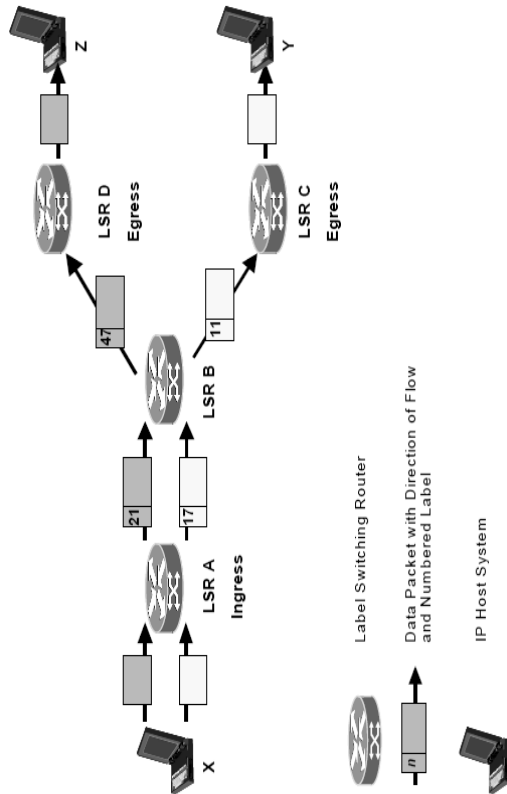


Figure 1 (Basic operation of MPLS)

MPLS uses a technique known as label switching to forward data through the network. A small, fixed-format label is inserted in front of each data packet on entry into the MPLS network. At each hop across the network, the packet is routed based on the value of the incoming label and dispatched to an outwards interface with a new label value. The path that data traverses through a network is defined by the transition in label values, as the label is swapped at each LSR. [3] Since the mapping between labels is constant at each LSR, the path is determined by the initial label value. Such a path is called a Label Switched Path (LSP). At the ingress to an MPLS network, each packet is examined to determine which LSP it should use and hence what label to assign to it. This decision is a local matter but is likely to be based on factors including the destination address, the quality of service requirements and the current state of the network. This flexibility is one of the key elements that make MPLS so useful. The set of all packets that are forwarded in the same way is known as a Forwarding Equivalence Class (FEC). One or more

FECs may be mapped to a single LSP. Figure 1 shows two data flows from host X: one to Y, and one to Z. Two LSPs are shown.

- LSR A is the ingress point into the MPLS network for data from host X. When it receives packets from X, LSR A determines the FEC for each packet, deduces the LSP to use and adds a label to the packet. LSR A then forwards the packet on the appropriate interface for the LSP.
- LSR B is an intermediate LSR in the MPLS network. It simply takes each labeled packet it receives and uses the pairing {incoming interface, label value} to decide the pairing {outgoing interface, label value} with which to forward the packet. This procedure can use a simple lookup table and, together with the swapping of label value and forwarding of the packet, can be performed in hardware. This allows MPLS networks to be built on existing label switching hardware such as ATM and Frame Relay. [4] This way of forwarding data packets is potentially much faster than examining the full packet header to decide the next hop. In the example, each packet with label value 21 will be dispatched out of the interface towards LSR D, bearing label value 47. Packets with label value 17 will be re-labeled with value 11 and sent towards LSR C.
- LSR C and LSR D act as egress LSRs from the MPLS network. These LSRs perform the same lookup as the intermediate LSRs, but the {outgoing interface, label value} pair marks the packet as exiting the LSP. The egress LSRs strip the labels from the packets and forward them using layer 3 routing. So, if LSR A identifies all packets for host Z with the upper LSP and labels them with value 21, they will be successfully forwarded through the network.

**Note** that the exact format of a label and how it is added to the packet depends on the layer 2 link technology used in the MPLS network. For example, a label could correspond to an ATM VPI/VCI, a Frame Relay DLCI, or a DWDM wavelength for optical networking. [5] For other layer 2 types (such as Ethernet and PPP) the label is added to the data packet in an MPLS “shim” header, which is placed between the layer 2 and layer 3 headers.

## 2. Label Distributions

In order that LSPs can be used, the forwarding tables at each LSR must be populated with the mappings from {incoming interface, label value} to {outgoing interface, label value}. This process is called LSP setup, or Label Distribution. [6][7][8] The MPLS architecture document (draft-ietf-mpls-arch) does not mandate a single protocol for the distribution of labels between LSRs. In fact it specifically allows for multiple protocols for use in different scenarios. Several different approaches to label distribution can be used depending on the requirements of the hardware that forms the MPLS network, and the administrative policies used on the network. The underlying principles are that an LSP is set up either in response to a request from the ingress LSR (downstream-on-demand), or preemptively by LSRs in the network, including the egress LSR (downstream unsolicited). It is possible for both to take place at once and for the LSP to meet in the middle. In all cases, labels are allocated from the downstream direction (where downstream refers to the direction of data flow, and this means that are advertised towards the data source). Thus, in the example in Fig.1, LSR D informs LSR B that LSR B should use label 47 on all packets for host Z. LSR B allocates a new label (21), enters the mapping in its forwarding table, and informs LSR A that it should use label 21 on all packets for host Z. Some possible options for controlling how LSPs are set up, and the protocols that can be used to achieve them, are described below.

- Hop-by-hop label assignment is the process by which the LSP setup requests are routed according to the next-hop routing towards the destination of the data. LSP setup could be initiated by updates to the routing table, or in response to a new traffic flow. The IETF MPLS Working Group has specified (but not mandated) LDP as a protocol for hop-by-hop label assignment. RSVP and CR-LDP can also be used.
- In Downstream Unsolicited label distribution, the egress LSR distributes the label to be used to reach a particular host. The trigger for this will usually be new routing information received at the egress node. Additionally, if the label distribution method is Ordered Control, each upstream LSR distributes a label further upstream. This effectively builds a tree of LSPs rooted at each egress LSR. LDP is currently the only protocol suitable for this mode of label distribution.
- Once LSPs have been established across the network, they can be used to support new

routes as they become available. As the routing protocols (for example BGP) distribute the new routing information upstream, they can also indicate which label (i.e. which LSP) should be used to reach the destinations to which the route refers.

- If an ingress LSR wants to set up an LSP that does not follow the next-hop routing path, it must use a label distribution protocol that allows specification of an Explicit Route. This requires downstream-on-demand label distribution. CR-LDP and RSVP are two protocols that provide this function.
- An ingress LSR may also want to set up an LSP that provides a particular level of service by, for example, reserving resources at each intermediate LSR along the path. In this case, the route of the LSP may be constrained by the availability of resources and the ability of nodes to fulfill the quality of service requirements. CR-LDP and RSVP are two protocols that allow downstream-on-demand label distribution to include requests for specific service guarantees. Figure 2 Shows MPLS label distribution process.

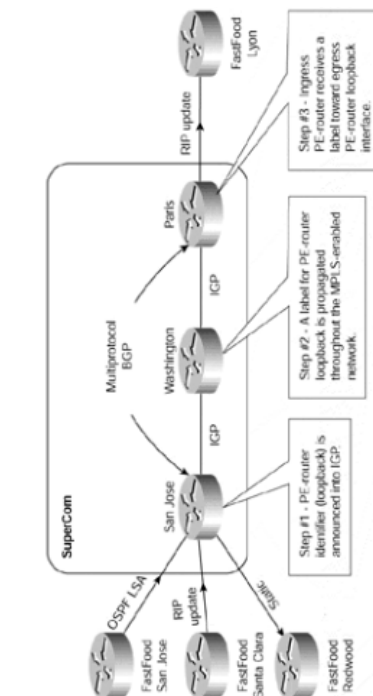


Figure 2 (Label Distribution)

### 3. Explicit Routers

An Explicit Route (ER) is most simply understood as a precise sequence of steps from ingress to egress. An LSP in MPLS can be set up to follow an explicit path, i.e. a list of IP addresses. However, it does not need to be specified this fully. For example, the route could specify only the first few hops. After the last explicitly specified hop has been reached, routing of the LSP proceeds using hop-by-hop routing. A component of an explicit route may also be less precisely specified. A collection of nodes, known as an Abstract Node, may be presented as a single step in the route, for example by using an IP prefix rather than a precise address. The LSP must be routed to some node within this Abstract Node as the next hop. The route may contain several hops within the Abstract Node before emerging to the next hop specified in the Explicit Route. An Explicit Route may also contain the identifier of an Autonomous System (AS). This allows the LSP to be routed through an area of the network that is out of the administrative control of the initiator of the LSP. The route may contain several hops within the Autonomous System before emerging to the next hop specified in the Explicit Route. An Explicit Route may be classified as "strict" or "loose". A strict route must contain only those nodes, Abstract Nodes or Autonomous Systems specified in the Explicit Route, and must use them in the order specified. A loose route must include all of the hops specified, and must maintain the order, but it may also include additional hops as necessary to reach the hops specified. Once a loose route has been established it can be modified (as a hop-by-hop route could be) or it can be "pinned" so that it does not change. [9][10] Explicit routing is particularly useful to force an LSP down a path that differs from the one offered by the routing protocol. It can be used to distribute traffic in a busy network, to route around network failures or hot spots, or to provide pre-allocated back-up LSPs to protect against network failures. Figure 3 shows MPLS explicit routes.

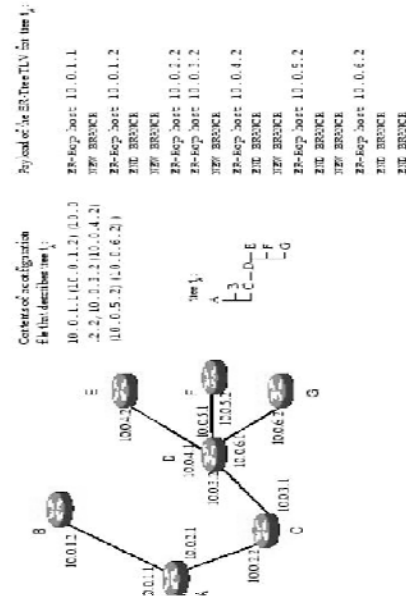


Figure 3 (Explicit Routes)

### 4. Constrained Routes

The route that an LSP may take can be constrained by many requirements selected at the ingress LSR. An Explicit Route is an example of a constrained route where the constraint is the order in which intermediate LSRs may be reached. Other constraints can be imposed by a description of the traffic that is to flow and may include bandwidth, delay, resource class and priority. One approach is for the ingress LSR to calculate the entire route based on the constraints and information that it has about the current state of the network. This leads it to produce an Explicit Route that satisfies the constraints. The other approach is a variation on hop-by-hop routing where, at each LSR, the next hop is calculated using information held at that LSR about local resource availability. The two approaches are combined if information about part of the route is unavailable (for example, it traverses an Autonomous System). In this case the route may be loosely specified in part, and explicitly routed using the constraints where necessary. Figure 4 shows MPLS constrained route.

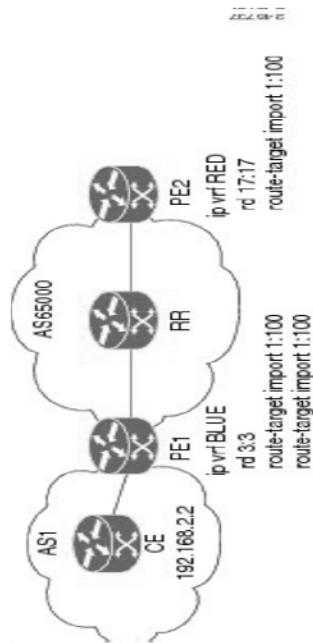


Figure 4 (Constrained Route)

### 5. Resource Reservation

In order to secure promised services, it is not sufficient simply to select a route that can provide the correct resources. These resources must be reserved to ensure that they are not shared or “stolen” by another LSP. The traffic requirements can be passed during LSP setup (as with constraint-based routing). They are used at each LSR to reserve the resources required, or to fail the setup if the resources are not available. Figure 5 shows MPLS resource reservation process.

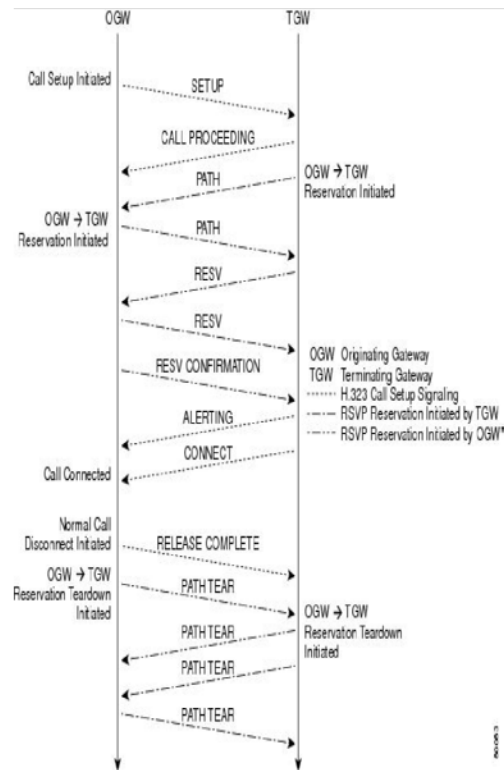


Figure 5 (Resource Reservation)

### 6. Traffic Engineering

Traffic Engineering is the process where data is routed through the network according to a management view of the availability of resources and the current and expected traffic. The class of service and quality of service required for the data can also be factored into this process. Traffic Engineering may be under the control of manual operators. They monitor the state of the network and route the traffic or provision additional resources to compensate for problems as they arise.[11][12][13] Alternatively, Traffic Engineering may be driven by automated processes reacting to information fed back through routing protocols or other means. Figure 6 below an extensive MPLS traffic engineering.

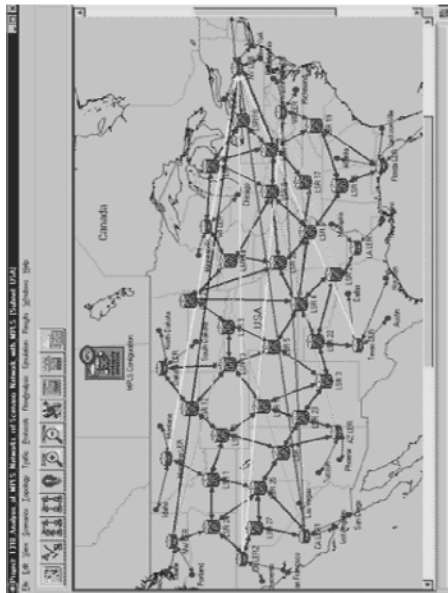


Figure 6 (Traffic Engineering)

## 7. Service Level Contracts

Many uses of the Internet require particular levels of service to be supplied. For example, voice traffic requires low delay and very small delay variation. Video traffic adds the requirement for high bandwidth. Customers increasingly demand service contracts that guarantee the performance and availability of the network.[14] In the past, in order to meet these requirements, network providers have had to over-provision their physical networks. MPLS offers a good way to avoid this issue by allocating the network resources to particular flows using constraint-based routing of LSPs.

## 8. Virtual Private Networks

A Virtual Private Network (VPN) allows a customer to extend their private network across a wider public network in a secure way. ISPs offer this service by ensuring that entry points to their network can exchange data only if they are configured as belonging to the same VPN. MPLS LSPs provide an excellent way to offer this service over an IP network.

## 9. Meeting the Needs of the Modern Network

VPNs have been addressed with additions to the BGP routing protocol, but IP has not provided good solutions to the requirements set out in the previous three sections. There has been no way of providing a guarantee of service, because the network is connectionless. Destination-based routing along shortest path routes tends to overload some links

and leave others unused. A popular solution is to use an overlay network, for example running IP over ATM PVCs. This is notoriously hard to manage, because many resources must be configured at each router in the network, and because there are two distinct protocols to be configured. It also leads to scaling issues, with an order of  $n^2$  connections needed in a network with  $n$  nodes. MPLS allows the use of just one set of protocols in the network. Using MPLS to meet the aims described in the previous three sections while avoiding the problems described above requires a label distribution protocol that supports Explicit Routes and constraint-based routing. There are currently two label distribution protocols that meet this definition: CR-LDP and RSVP. There is a debate about which of these protocols is preferable, which is most suitable for particular scenarios, and whether it is necessary to implement both of the protocols in an MPLS network. Since the LSPs set up to support Traffic Engineering, Service Contracts and VPNs are all configured in the same way for RSVP and CR-LDP (through the Traffic Engineering MIB), they are referred to as Traffic Engineered LSPs.

## 10. Conclusion

Link failure is a common cause of service disruption in computer networks. Many techniques have been developed to alleviate the consequences of hardware failure in a network like the Internet by rerouting traffic from a failed link to a working or a set of working links. Rerouting is performed automatically in the Internet by recomputing routing tables. However routing convergence may be slow and faster techniques which require expensive hardware have been developed to protect networks from link failures. MPLS is a recent virtual circuit packet switching technology which has been designed to support the forwarding of IP packets over virtual circuits. MPLS Fast Reroute is a traffic engineering technique that is able to reroute IP traffic quickly without the need of additional hardware. Indeed, MPLS Fast Reroute relies on pre-planned backup path to reroute traffic on a link failure and can be implemented in existing routers. An important delivery mode of the Internet is multicasting, where the information sent by a member of a multicast group is received by all other members of the group. A popular example of a multicasting application is teleconferencing. In real-time applications like teleconferencing, if a link failure occurs, it is crucial to repair the multicast routing tree of the multicast communication in a short time. For example, an interruption of service of more than 50 ms is noticeable in a live transmission. Establishing a backup path to protect a multicast routing tree is a resource consuming process. Therefore, it is

desirable to protect a large number of members of a multicast group with a low number of backup paths. In this thesis, we presented an algorithm which is able to choose such a backup path, and the design and implementation of an MPLS-based rerouting mechanism adapted to the protection of multicast routing trees. We now review our contributions and expose possible future work.

## References

- [1] Cisco Systems, Inc. (2004). *Managed VPN – Van Wijnen and Versatel*. Retrieved November 19th, 2007, from [http://www.cisco.com/en/US/netsol/ns465/networking\\_solutions\\_customer\\_profile0900aecd801aa3f5.html](http://www.cisco.com/en/US/netsol/ns465/networking_solutions_customer_profile0900aecd801aa3f5.html)
- [2] Cisco Systems, Inc. (2005). *From Frame Relay to IP VPN: Why to Migrate, Why to Out-Task*. Retrieved November 18th, 2007, from [http://www.cisco.com/en/US/netsol/ns458/networking\\_solutions\\_white\\_paper0900aecd8017a894.shtml](http://www.cisco.com/en/US/netsol/ns458/networking_solutions_white_paper0900aecd8017a894.shtml)
- [3] Layer 2 MPLS VPN. (2007, October 20). In *Wikipedia, The Free Encyclopedia*. Retrieved 18:03, November 18, 2007, from [http://en.wikipedia.org/w/index.php?title=Layer\\_2\\_MPLS\\_VPN&oldid=165912463](http://en.wikipedia.org/w/index.php?title=Layer_2_MPLS_VPN&oldid=165912463)
- [4] Cisco Systems, Inc. (2006, June). *Understand MPLS Technology. MPLS TE Technology Overview*. (chap. 2). Retrieved November 19th, 2007, from <http://downloads.techrepublic.com.com/thankyou.aspx?authId=uqqOzCBTkk7ekSZjOPwgf9Z5C6ZJWYXLNMI0MVnKEACzi6IN9H2AHB5e56BBkXn&q=MPLS%20T>
- [9] Juniper Networks. *Traffic Engineering for the New Public Network*. [http://www.omimo.be/magazine/00q4/2000q4\\_p054.pdf](http://www.omimo.be/magazine/00q4/2000q4_p054.pdf)
- [10] Sprint Nextel, Inc. (2006, January). *Sprint Global MPLS VPN IP Whitepaper*. Retrieved November 19th, 2007, from <http://whitepapers.techrepublic.com.com/thankyou.aspx?authId=uqqOzCBTkk7ekSZjOPwgf9Z5C6ZJWYXLNMI0MVnKEADJ90SewjXUM22n4A2PUWMB&&q=Sprint+Global+MPLS+VPN&docid=273906&view=273906&load=1>
- [11] Verizon, Inc. (2006, December). *MPLS VPN Networking and Migration Considerations*. Retrieved November 18th, 2007, from <http://whitepapers.techrepublic.com.com/thankyou.aspx?&q=MPLS+VPN+Networking+and+Migration+Verizon&docid=284829&view=284829>
- E%20overview%20cisco&docid=177738&view=177738&load=1
- [5] AT&T Knowledge Ventures. (2007, July 25). *Transitioning to an MPLS Network*. Retrieved November 19th, 2007, from [http://www.business.att.com/nx\\_resource.jsp?repid=Topic&rtype=Whitepaper&rvalue=eb\\_fpoc\\_navigating\\_to\\_mpls\\_enabled\\_networks&repoitem=vpns&segment=ent\\_biz](http://www.business.att.com/nx_resource.jsp?repid=Topic&rtype=Whitepaper&rvalue=eb_fpoc_navigating_to_mpls_enabled_networks&repoitem=vpns&segment=ent_biz) <http://www.ietf.org/html.charters/rsvp-charter.html> <http://www.ietf.org/html.charters/mpls-charter.html>
- [6] AT&T Knowledge Ventures. (2007, August 31). *Understanding VPN Technology Choices: Comparing MPLS, IPSec and SSL*. Retrieved November 19th, 2007, from [http://www.business.att.com/nx\\_resource.jsp?repid=Topic&rtype=Whitepaper&rvalue=understanding\\_vpn\\_technology\\_choices&repoitem=vpns&segment=ent\\_biz&guid=4BFDAE84-C61B-416F-886A-F606E9678B1C;08905D72-1FE7-450C-8EA5-B5F1565DD558](http://www.business.att.com/nx_resource.jsp?repid=Topic&rtype=Whitepaper&rvalue=understanding_vpn_technology_choices&repoitem=vpns&segment=ent_biz&guid=4BFDAE84-C61B-416F-886A-F606E9678B1C;08905D72-1FE7-450C-8EA5-B5F1565DD558)
- [7] Cisco Systems, Inc. (2004). *Managed VPN – Analysis and Comparisons of MPLS-Based IP VPN Security*. Retrieved November 18th, 2007, from [http://www.cisco.com/en/US/netsol/ns465/networking\\_solutions\\_white\\_paper09186a008020c5a6.shtml](http://www.cisco.com/en/US/netsol/ns465/networking_solutions_white_paper09186a008020c5a6.shtml)
- [8] Cisco Systems, Inc. (2004). *Managed VPN – Comparison of MPLS, IPSec, and SSL Architecture – Comparing MPLS, IPSec, and SSL*. Retrieved November 19th, 2007, from [http://www.cisco.com/en/US/netsol/ns465/networking\\_solutions\\_white\\_paper0900aecd801b1b0f.shtml](http://www.cisco.com/en/US/netsol/ns465/networking_solutions_white_paper0900aecd801b1b0f.shtml)
- [12] Martini draft. (2007, April 2). In *Wikipedia, The Free Encyclopedia*. Retrieved 18:03, November 18, 2007, from [http://en.wikipedia.org/w/index.php?title=Martini\\_draft&oldid=119746107](http://en.wikipedia.org/w/index.php?title=Martini_draft&oldid=119746107)
- [13] Multiprotocol Label Switching. (2007, November 7). In *Wikipedia, The Free Encyclopedia*. Retrieved 18:04, November 18, 2007, from [http://en.wikipedia.org/w/index.php?title=Multiprotocol\\_Label\\_Switching&oldid=169803565](http://en.wikipedia.org/w/index.php?title=Multiprotocol_Label_Switching&oldid=169803565)
- [14] Pseudo-wire. (2007, November 17). In *Wikipedia, The Free Encyclopedia*. Retrieved 18:01, November 18, 2007, from <http://en.wikipedia.org/w/index.php?title=Pseudowire&oldid=172121304>

# Block Based Video Watermarking Scheme Using Wavelet Transform and Principle Component Analysis

Nisreen I. Yassin<sup>1</sup>, Nancy M. Salem<sup>2</sup>, and Mohamed I. El Adawy<sup>3</sup>

<sup>1</sup> National Research Centre, Cairo, Egypt.

<sup>2</sup> Department of Biomedical Engineering, Faculty of Engineering, Helwan University, Egypt.

<sup>3</sup> Department of Comm., Elect., and Computers, Faculty of Engineering, Helwan University, Egypt.

## Abstract

In this paper, a comprehensive approach for digital video watermarking is introduced, where a binary watermark image is embedded into the video frames. Each video frame is decomposed into sub-images using 2 level discrete wavelet transform then the Principle Component Analysis (PCA) transformation is applied for each block in the two bands LL and HH. The watermark is embedded into the maximum coefficient of the PCA block of the two bands. The proposed scheme is tested using a number of video sequences. Experimental results show high imperceptibility where there is no noticeable difference between the watermarked video frames and the original frames. The computed PSNR achieves high score which is 44.097 db. The proposed scheme shows high robustness against several attacks such as JPEG coding, Gaussian noise addition, histogram equalization, gamma correction, and contrast adjustment.

**Keywords:** Digital video watermarking, Principal Component Analysis, Discrete Wavelet Transform, Binary watermark.

## 1. Introduction

Digital watermarking is a new technology used for copyright protection of digital media. Digital watermarking was introduced at the end of the 20<sup>th</sup> century to provide means of enforcing copyright protection of digital data. Where, ownership information data called watermark is embedded into the digital media (image, audio, and video) without affecting its perceptual quality. In case of any dispute, the watermark data can be detected or extracted from the media and used as a proof of ownership. Imperceptibility and robustness against attacks are the fundamental issues in digital watermarking techniques [1-2]. Recently, digital video watermarking has emerged as a significant field of interest and a very active area of research [3]. Many digital watermarking schemes have been proposed for video. Most these schemes are based on the techniques of image watermarking, but video watermarking has some issues not present in image watermarking. This is because video sequences have some distinguish characteristics such as the temporal and inter-frame characteristics, which require specific approaches for

video watermarking [4-7]. Video watermarking schemes can be classified into two main categories based on the domain which used for hiding the watermark bits in the host video. The first one is the spatial domain watermarking where embedding and detection of watermark is performed by directly manipulating the pixel intensity values of the video frame [8-9]. The second category is the transform domain techniques [10-12] in which the watermark is embedded by changing the frequency components. The commonly used transform domain techniques are Discrete Fourier Transform (DFT), the Discrete Cosine Transform (DCT), the Discrete Wavelet Transform (DWT), and Principle Component Analysis transform. The frequency domain watermarking schemes are relatively more robust than the spatial domain watermarking schemes, particularly in lossy compression, noise addition, pixel removal, rescaling, rotation and cropping.

Swanson [13] has proposed a scene-based video watermarking procedure in which the watermark is generated from a temporal wavelet transform of the video scenes. In Inoue [14] the watermark was embedded in the lowest frequency components of each frame in the uncoded video using a controlled quantization process. Chan *et al.* [15] propose a hybrid digital video watermarking scheme based on the scene change analysis and error correction code. He has used the Discrete Wavelet Transform by embedding in frequency coefficients of video frames. Hussein [16] embeds the watermark data to the HL and LH bands of the wavelet domain using motion estimation approach. The motion in these bands does not affect the quality of the frame.

The PCA domain was first introduced to gray-scale image watermarking by Thai D. Hien *et al.* [17]. In [18] the PCA transform is used to embed the watermark in each RGB color channel of each frame of the video, where the same or multi-watermark can be embedded into the three color channels of the image in order to increase the robustness of the watermark. The main



advantage of using PCA transform is to choose the suitable significant components into which to embed the watermark. In Yavuz [19], a reference image is generated from the cover image using PCA and the watermark is embedded according to the difference between the image and its reference. Kang [20] has proposed a new algorithm where advantage of the strength of both multi-band wavelet transform (MWT) and PCA is used. The watermark energy is distributed to wavelet coefficients of every detail sub-band efficiently to achieve better robustness and perceptual transparency. A hybrid scheme combining both DWT and PCA has been proposed by Mostafa *et al.* in [21]. The watermark was embedded into the first principle components and the mid-band coefficient of the PCA wavelet frame. In Sinha [22] a binary watermark is embedded into each of the video frames by the decomposition of the frames into DWT sub bands followed by block based PCA on the sub-blocks of the low frequency sub-band. The watermark is embedded into the principal components of the sub-blocks.

In this paper, we propose an imperceptible and robust video watermarking algorithm based on DWT and PCA. DWT is more computationally efficient than other transform methods because of its excellent localization properties which provide the compatibility with the Human Visual System (HVS). This paper is organized as follows: section 2 presents the proposed watermarking scheme. Section 3 introduces the experimental results and the conclusion is given in section 4.

## 2. Proposed Watermarking Scheme

The proposed hybrid watermarking scheme is based on the combination of DWT and PCA.

### 2.1 Discrete Wavelet Transform

The DWT is used in a wide variety of signal processing applications [23]. 2-D discrete wavelet transform (DWT) decomposes an image or a video frame into sub-images, 3 details and 1 approximation. The approximation sub image is lower resolution approximation image (LL) however the details sub images are horizontal (HL), vertical (LH) and diagonal (HH) detail components. The process can then be repeated to compute multiple "scale" wavelet decomposition. The main advantage of the wavelet transform is its compatibility with a model aspect of the HVS as compared to the FFT or DCT. This allows us to use higher energy watermarks in regions that the HVS is known to be less sensitive, such as the high resolution detail bands. Embedding watermarks in these regions allow us to increase the robustness of our watermark without any visible impact on the image quality. In the proposed algorithm, sub-bands LL and HH from resolution level 2 of the wavelet transform of the frame

are chosen for the embedding process. The following figure shows the selected DWT bands which used in our proposed algorithm.

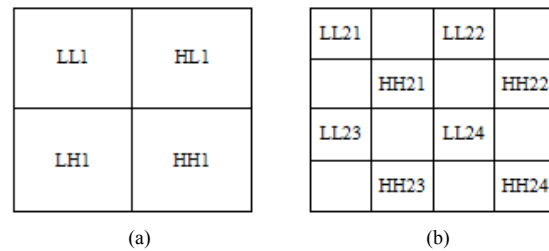


Fig.1 DWT sub-bands in (a) level 1, (b) level 2.

Embedding the watermark in low frequencies obtained by wavelet decomposition increases the robustness against attacks like filtering, lossy compression and geometric distortions while making the scheme more sensitive to contrast adjustment, gamma correction, and histogram equalization. Embedding the watermark in high frequency sub-bands makes the watermark more imperceptible while embedding in low frequencies makes it more robust against a variety of attacks.

### 2.2 Principal Component Analysis

Principal component analysis (PCA) is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of correlated variables into a set of values of uncorrelated variables called principal components. PCA plots the data into a new coordinate system where the data with maximum covariance are plotted together and is known as the first principal component. Similarly, there are the second and third principal components and so on. The first principal component has the maximum energy concentration [24].

The diagram in Fig. 2 shows the embedding and extraction process of the watermark. In the proposed scheme, a binary image is embedded in the LL DWT sub-bands of level 2 of each decomposed frame in the video. Also, the same binary image is embedded in the HH DWT sub-band of level 2 of each decomposed frame. Embedding the watermark in both LL and HH makes the scheme robust to a variety of low and high frequency characteristic attacks. The extraction procedure of the watermark is similar to the embedding one.

### 2.3 Embedding Procedure

Step 1: Divide video into frames ( $2N \times 2N$ ), then convert RGB frames to YUV frames.

Step 2: Choose the luminance component Y of each frame and apply DWT on it. This result in four multi-resolution sub-bands ( $N \times N$ ): LL1, HL1, LH1, and HH1. For each band apply DWT again to get 16 sub-bands ( $N/2 \times N/2$ ). From these sub-bands, select (LL21, LL22, LL23, LL24, HH21, HH22, HH23, and HH24).

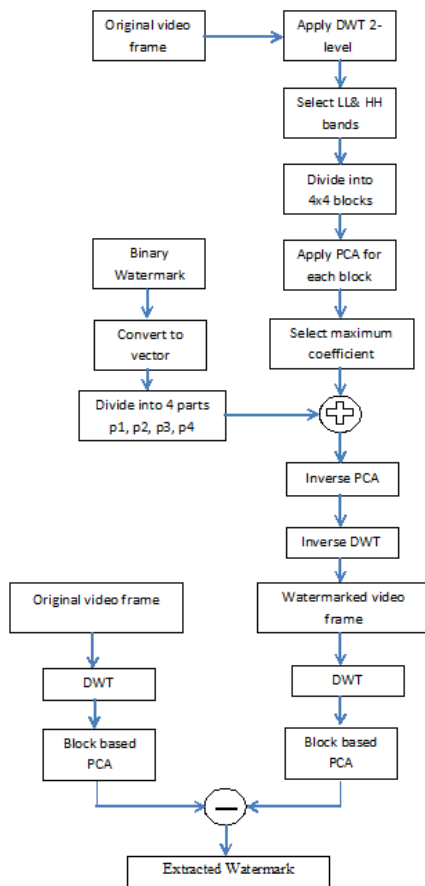


Fig. 2 Watermark embedding and extraction algorithm.

Step 3: Divide each selected sub-bands  $I_s$  with  $N/2 \times N/2$  dimension into  $n \times n$  non-overlapping blocks where the number of blocks is  $k = (N/2 \times N/2) / (n \times n)$ . Then, apply PCA to each block as described.

1. For each block  $B_{si}$  ( $n \times n$ ) compute the mean of the block  $m_i$ , where  $B_{si}$  represent block number  $i$  in the selected sub-band  $I_s$ . Then get the block zero mean  $A_i$  as follows:

$$A_i = E(B_{si} - m_i) \quad (1)$$

2. For each block, calculate the covariance matrix  $C_i$  of the zero mean block  $A_i$  as:

$$C_i = A_i \times A_i^T \quad (2)$$

Where  $T$  denotes the matrix transpose operation.

3. Transform each block into PCA components by calculating the eigenvectors corresponding to eigenvalues of the covariance matrix:

$$C_i \phi = \lambda_i \phi \quad (3)$$

Where  $\phi$  is the matrix of eigenvectors and  $\lambda$  is the matrix of eigenvalues.

4. Compute the PCA transformation of each block to get a block of uncorrelated coefficients by:

$$Y_i = \phi^T A_i \quad (4)$$

Where  $Y_i$  is the principle component of the  $i^{\text{th}}$  block.

Step 4: Convert the RGB  $32 \times 32$  watermark image to binary image. Convert the binary image into a vector  $W = \{w_1, w_2, \dots, w_{32 \times 32}\}$  of zeros and ones.

Step 5: Divide the vector  $W$  into four parts  $p_1, p_2, p_3$ , and  $p_4$ . Then  $p_1$  is embedded into each of the corresponding LL21 and HH21,  $p_2$  is embedded into each LL22 and HH22,  $p_3$  is embedded into each LL23 and HH23, and  $p_4$  is embedded into each LL24 and HH24. The watermark bits are embedded with strength  $\alpha$  into the maximum coefficient  $M_i$  of each PC block  $Y_i$ . The embedding equation is:

$$M_i = M_i \pm \alpha W \quad (5)$$

Where,  $\alpha$  is the watermark embedding strength. The value of  $\alpha$  in this algorithm is 9 for all selected wavelet bands. If the watermark bit is 1 then adding  $\alpha$  to the maximum coefficient in the  $Y$  block but if it is zero, then  $\alpha$  is subtracted from the same coefficient.

Step 6: Apply inverse PCA on the modified PC block  $Y_i$  to obtain the modified wavelet block by using:

$$A_i = \phi Y_i \quad (6)$$

Step 7: Apply the inverse DWT to obtain the watermarked luminance component of the frame. Finally reconstruct the RGB watermarked frame and obtain the watermarked video.

## 2.4 Watermark Extraction

The steps used for watermark extraction is the same as the steps in the embedding but in the reverse direction. The original video sequence is required for the extraction procedure so the algorithm is non-blind.

Step 1: Convert the watermarked video into frames. Each RGB frame is converted to YUV representation.

Step 2: For each  $Y$  component, apply DWT to decompose  $Y$  into 16 multi-resolution sub-bands. Choose LL and HH sub-bands and divide them into  $n \times n$  non-overlapping blocks.

Step 3: For each block, apply PCA transformation as described in the embedding procedure.

Step 4: Extract the watermark by applying the following equation:

$$W = \frac{M_i - M_i}{\alpha} \quad (7)$$

Step 5: The extracted watermark is compared with the original watermark by computing the similarity measure between them as follows:

$$NC = \frac{\sum_i \sum_j W(i,j) \cdot W'(i,j)}{\sum_i \sum_j W(i,j)^2} \quad (8)$$

Where, NC is the normalized correlation.

NC value is 1 when the watermark and the extracted watermark are identical and zero if the two are different from each other.

### 3. Experimental Results

A number of video sequences are used for testing the proposed scheme for example the foreman video sequence [12]. For evaluating the performance of any watermarking system, Peak Signal to Noise Ratio (PSNR) is used as a common measure of the visual quality of the watermarking system. To calculate the PSNR, first the Mean Square Error (MSE) between the original and watermarked frame is computed as follows:

$$MSE = \frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N [I(i,j) - I'(i,j)]^2 \quad (9)$$

Where M, N are the size of the frame, and I(i, j), I'(i, j) are the pixel values at location (i, j) of the original and watermarked frames. Then, PSNR is defined as:

$$PSNR = 10 \log_{10} \frac{255^2}{MSE} \quad (10)$$

The luminance component of the first 100 frames of the foreman video sequence are watermarked. The frame size is 256x256. The watermark is a binary image with size 32x32.

The original sampled frame and its corresponding watermarked frame are shown in Fig. 3. The measured PSNR is 44.0975 db and the watermarked frame appears visually identical to the original. The value of PSNR is constant over all the tested frames which means that the error between the original and watermarked frames is very low so high visual quality is obtained. Fig. 4 shows the original watermark and the extracted watermark from LL band and HH band where no attacks were applied. The measured value of NC is 1 for both LL band and HH band, i.e. the extracted watermark is identical to the original and exact extraction is obtained.



Fig. 3 (a) Original frame, (b) Watermarked frame (PSNR = 44.0975 db).

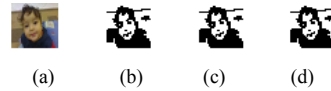


Fig. 4 (a) Original watermark, (b) Binary watermark, (c) Extracted LL watermark (NC=1), (d) Extracted HH watermark (NC=1).

To measure the robustness of our proposed scheme, the watermarked frame was subjected to a variety of attacks such as gamma correction, contrast adjustment, histogram equalization, and jpeg compression.

Fig. 5, illustrates the attacked frame by Gamma correction at different values 0.5, 2, and 4. The extracted watermarks from LL and HH are also illustrated. It is shown that, by the proposed algorithm the watermark can be easily extracted and recognized from both LL band and HH band.

Robustness against histogram equalization and Gaussian noise attacks are shown in Fig. 6. When adding Gaussian noise with zero mean and variance = 0.001 to the watermarked frame, the extracted watermark for both LL and HH bands is shown in Fig. 6. These watermarks can be easily recognized by human eyes. While Fig. 7 illustrates the performance of the proposed scheme in case of contrast adjustment attacks at factors 10 and 30.

Attack	Gamma correction 0.5	Gamma correction 2	Gamma correction 4
Attacked Frame			
	PSNR = 15.426 db	PSNR = 16.481 db	PSNR = 12.311 db
Extracted Watermark	 NC = 0.935 NC = 0.980	 NC = 0.584 NC = 0.820	 NC = 0.357 NC = 0.507

Fig. 5 Gamma correction attack.

Attack	Automatic Equalization	Gaussian noise 0.001
Attacked Frame		
	PSNR = 18.608 db	PSNR = 29.807 db
Extracted Watermark	 NC = 0.990 NC = 0.996	 NC = 0.880 NC = 0.859

Fig. 6 Automatic equalization and adding Gaussian noise attack.







Attack	Contrast (factor = 10)	Contrast (factor = 30)
Attacked Frame	 PSNR = 33.675 db	 PSNR = 22.881 db
Extracted Watermark	 LL NC = 1  HH NC = 1	 LL NC = 0.976  HH NC = 0.974

Fig. 7 Contrast adjustment attack.










JPEG 80%	JPEG 50%	JPEG 40%
 PSNR = 43.309	 PSNR = 42.546	 PSNR = 41.486
 LL NC = 0.953  HH NC = 1	 LL NC = 0.757  HH NC = 0.720	 LL NC = 0.675  HH NC = 0.564

Fig. 8 JPEG attacks.

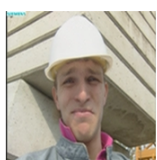

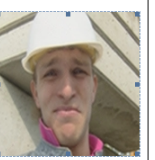






Attack	Resize 256 :512:256	Rotate 5° (Matlab)	Cropping (Matlab)
Attacked Frame	 PSNR = 46.161	 PSNR = 9.163	 PSNR = 15.889
Extracted Watermark	 LL NC = 1  HH NC = 1	 LL NC = 0.121  HH NC = 0.109	 LL NC = 0.070  HH NC = 0.062

Fig. 9 Some geometric attacks.

The watermark must be robust against JPEG compression attack since it is the common compression technique used in image compression. Fig. 8 shows the performance of the proposed scheme when subjected to such attack. As shown in Fig. 8, the value of NC is decreasing as the quality factor of JPEG decrease. At QF = 80%, the NC for LL and HH are 0.953 and 1 respectively. Also, At QF = 40% the NC for LL and HH are 0.675 and 0.564 respectively. In case of jpeg compression, the watermark can survive under quality factor 40%.

In case of geometric attacks, we test the scheme against frame resizing, frame rotation, and frame cropping. The

frame is resized from 256 x256 to 512x512 then back to its original size. The results show that the watermark is totally recovered as shown in Fig. 9. In same way rotate/resize and crop/resize of the frame gives the results shown in Fig. 9. From the results, the scheme is not robust against frame rotation and frame cropping which will be investigated in the future.

By comparing the proposed method with previous methods such as Mostafa [21] and Wang [25], we find that the proposed scheme registered a constant PSNR equal to 44.0975 db which is greater than the PSNR reported by both Mostafa and Wang as shown in table 1.

Table 1: PSNR Comparison

Algorithm	PSNR (db)
Mostafa [21]	39.0693
Wang [25]	32
Proposed	44.0975

Bit Error Rate (BER) is used as another performance metric to compare the performance of the proposed scheme. BER is the ratio of the number of bits recovered in error to the total number of received bits [26]. Table 2 contains some values of BER measured by Mostafa's method, Wang's method, and our proposed method.

Table 2: BER Comparison

Attack	Mostafa [21]	Wang [25]	Proposed Method
Salt & Pepper Noise (0.05)	46%	5%	37.7%
Gaussian noise (0.005)	42%	5%	40.6%
Sharpening	2%	17%	1.5%
Rotate 0.3°	20%	25%	49%
Smoothing	3%	25%	22%

## 4. Conclusions

A video watermarking scheme has been proposed in this paper. The algorithm is implemented using 2-level DWT in conjunction with PCA transform. This scheme is imperceptible and robust against several attacks. A binary watermark has been embedded into LL and HH bands of level 2 of DWT block based PCA. The proposed scheme has a good performance compared with previous schemes. As a future work, embedding the watermark into higher levels of the wavelet transform will be investigated. Collecting other transformations together to enhancement the performance of the proposed scheme against geometric attacks will be studied.

## References

- [1] M. K. Thakur, V. Saxena, and J. P. Gupta, "A Performance analysis of objective video quality metrics for digital video

watermarking”, 3<sup>rd</sup> IEEE International Conference on Computer Science and Information Technology – ICCSIT '10, 9-11 July, 2010, pp.12-17, Chengdu, China.

[2] S. Voloshynovskiy, S. Pereira, and T. Pun, “Watermark attacks”, Erlangen Watermarking Workshop 99, October 1999.

[3] C.I. Podilchuk and E.J. Delp, “Digital watermarking: algorithms and applications”, IEEE Signal Processing Magazine, Vol. 18, Issue 4, July 2001, pp. 33-46.

[4] P.W. Chan, M.R. Lyu, and R.T. Chin, “A Novel scheme for hybrid digital video watermarking”, IEEE Transactions on Circuits and Systems For Video Technology, Vol. 15, No. 12, December 2005.

[5] G. Doërr and J.L. Dugelay, “A guide tour of video watermarking”, Signal Processing: Image Commun., April 2003, Vol. 18, No. 4, pp. 263–282.

[6] Y. R. Lin, H.Y. Huang and W.H Hsu, “An embedded watermark technique in video for copyright protection”, 18<sup>th</sup> International Conference on Pattern Recognition – ICPR '06, 20-24 August 2006, pp. 795- 798, Hong Kong.

[7] C.V. Serdean, M.A. Ambroze., M. Tomlinson, and J.G. Wade, “DWT based video watermarking for copyright protection, invariant to geometrical attacks”, IEE on Vision, Image and Signal Processing, Vol. 150, Issue 1, 2003, pp. 51-58.

[8] R. Chandramouli and N. Memon, “Analysis of LSB based image steganography techniques”, in Proceedings International Conference on Image Processing, 7-10 October, 2001, Vol. 3, pp. 1019–1022, Thessaloniki, Greece.

[9] G. Langelaar, I. Setyawan, and R. Lagendijk, “Watermarking digital image and video data”, IEEE Signal Processing Magazine, Vol. 17, No. 9, September 2000, pp. 20–43.

[10] I. J. Cox, J. Kilian, F. T. Leighton and T. Shamoon, “Secure spread spectrum watermarking for multimedia”, IEEE Transactions on Image Processing, Vol. 6, Issue 12, 1997, pp. 1673-1687.

[11] C.H. Li and S.S. Wang, “Transform-based watermarking for digital images and video”, International Conference on Consumer Electronics –ICCE '99, 22-24 June, 1999, Los Angeles, USA.

[12] S. Sinha, S. Pramanick, A. Jagatramka, P. Bardhan, D. K. Kole, and A. Chakraborty, “Digital video watermarking using singular value decomposition”, Proceedings on IEEE EDS Student Paper Conference, 2011, pp. 53-56.

[13] M. D. Swanson, B. Zhu, and A. H. Tewfik, “Multiresolution scene-based video watermarking using perceptual models”, IEEE Journal on Selected Areas in Communications, Vol. 16, No. 4, May 1998.

[14] H. Inoue, A. Miyazaki, T. Araki, and T. Katsura, “A digital watermark method using the wavelet transform for video data”, Proceedings of the 1999 IEEE International Symposium on Circuits and Systems – ISCAS '99, Vol. 4, 1999, pp. 247-250.

[15] P.W. Chan and M. R. Lyu, “A DWT-based digital video watermarking scheme with error correcting code” Proceedings of the 5<sup>th</sup> International Conference on Information and Communications Security, 2003, pp. 202-213.

[16] J. Hussein and A. Mohammed, "Robust video watermarking using multi-band wavelet transform", International Journal of Computer Science Issues, IJCSI, Vol. 6, Issue 1, November 2009, pp. 44-49.

[17] T. D. Hien, Y.W. Chen, and Z. Nakao, “A robust digital watermarking technique based on principal component analysis” International Journal of Computational Intelligence and Applications, Vol. 4, No. 2, 2004, pp. 138-192.

[18] H. Mirza, H. Thai, and Z. Nakao, “Digital video watermarking based on RGB color channels and principal component analysis”, Lecture Notes in Computer Science, Vol. 5178, 2008, pp. 125-132, Springer-Verlag Berlin Heidelberg.

[19] E. Yavuz and Z. Telatar, “Digital watermarking with PCA based reference Images”, Lecture Notes in Computer Science, Vol. 4678, 2007, pp.1014-1023, Springer-Verlag Berlin Heidelberg.

[20] X. Kang ,W. Zeng ,and J. Huang, “A Multi-band wavelet watermarking scheme”, International Journal of Network Security, Vol. 6 , No. 2, March 2008, pp. 121–126.

[21] S. A. Mostafa, A. S. Tolba, F. M. Abdelkader, and H. M. Elhindy, “Video watermarking scheme based on principal component analysis and wavelet transform”, International Journal of Computer Science and Network Security, Vol. 9, No. 8, August 2009.

[22] S. Sinha, P. Bardhan, S. Pramanick, A. Jagatramka, D. K. Kole, and A. Chakraborty, “Digital video watermarking using discrete wavelet transform and principal component analysis”, International Journal of Wisdom Based Computing, Vol. 1, No. 2, August 2011.

[23] H. Olkkonen, “Discrete wavelet transforms - algorithms and applications”, InTech, August 2011.

[24] R. C. Gonzalez, R. E. Woods, and S. L. Eddins, “Digital image processing Using Matlab”, Pearson Prentice Hall, New Jersey, 2004.

[25] Hao-Xian Wang, J. Zhe-Ming Lu and Sheng-He Sun, “A modified video watermarking algorithm based on SVD in the DCT domain”, International Journal of Computer Sciences and Engineering system, Vol. 2, No. 1, January 2008, pp. 37-40.

[26] A.T.S. Ho, Y.Q. Shi, H.J. Kim, and M. Barni, “Digital Watermarking”, 8<sup>th</sup> International Workshop on Digital Watermarking IWDW '9, Lecture Notes in Computer Science, Vol. 5703, August 2009, Guildford, UK.

**Nisreen I. Yassin** received the B.Sc. and M.Sc. degrees in communications engineering from Faculty of Engineering, Helwan University, Cairo, Egypt in 2001 and 2005, respectively. Currently she is pursuing a Ph.D. degree in the Department of Electronics, Communications, and Computers Engineering, Faculty of Engineering, Helwan University, Egypt. She is working as associate researcher in the National Research Center, Cairo, Egypt. Her research interests include image processing and digital watermarking.

**Nancy M. Saleem** graduated with a BSc degree in communications and electronics engineering from Helwan University, Cairo, Egypt, in 1998 and an MSc in 2003. She received her PhD degree from the Department of Electrical Engineering and Electronics, University of Liverpool, UK in 2007. Currently she is working as a lecturer in Department of Biomedical Engineering, Faculty of Engineering, Helwan University. Her research interests include digital image processing, medical imaging, machine learning and clustering algorithms.

**Mohamed I. Eladawy** graduated from the Department of Electrical Engineering, Faculty of Engineering of Assiut University in May 1974; M.Sc. from Cairo University in May 1979; Ph.D. from Connecticut State University, School of Engineering, in May 1984. He worked as an Instructor at the Faculty of Engineering, Helwan University since 1974. Currently he is a Professor at the Department of Communication and Electronics Engineering and the Vice Dean for Student Affairs in the same faculty. He was working for the general organization for technical and vocational training for 6 years from 1989 to 1995 in Saudi Arabia. His research interests are in signal processing and its medical applications.

# Tools for decision support in planning academic needs of actors

LATIFA. Oubedda<sup>1</sup>, BRAHIM. Erraha<sup>1</sup>, MOHAMED. Khalfaoui<sup>2</sup>

(1) Laboratory of Industrial Engineering and Computer Science (LG2I), National School of Applied Sciences -Agadir, ,  
University Ibn Zohr

(2) Analysis Laboratory for Systems, Information Processing and Integrated Management (lastima), Ecole Supérieure de  
Technologie de Salé, University Mohammed V Agdal

## Abstract

The objective of this work is to establish an information system which would facilitate decision making for the exploitation of a model consisting of the main university stakeholders (teachers, students and administrators). This system is based on the relationship between actors (players) on the one hand and their activities and their aggregations in a graduate level on the other. It aims to make available to managers of the university a set of dashboards that can improve the quality of education.

We will start by modeling the actors upstream and we will study the processes on their own organizations, their activities and their aggregations. This approach is based on the analysis made by the actors to switch to an information technology approach in the process of searching for knowledge. The first applications of this work focus on data related to the department of English Studies at the Faculty of Arts and Humanities at Ibn Zohr University in Agadir, Morocco. The results are encouraging and can be generalized to all courses offered by academic institutions.

**Keywords:** Information System, Decision Support Information System, data warehouse, databases, multidimensional analysis.

## 1. Introduction

Business Intelligence is one of the areas of computer science that is experiencing a boom today. Indeed, the managers of sectors which are facing more and more unstable environments are expected to take more effective decisions based on reliable data. The challenge now is not only to have a better tool for decision making, but it is to spot that input data which is not effective. Thus, the design of information systems of decision which are to be tailored and scalable is a timely issue for organizations all over the world.

Today, the missions assigned to the university and the organization and powers vested in it by the 01-00 law require it to be not only to the needs expressed by its users, but also to anticipate these needs by acting as a University-Business. Furthermore, business intelligence has recently become a necessity for the systematic control of Moroccan universities.

This work is therefore bound to this context and focuses on the design and implementation of an information system specialized in refining the university environment [7]. The goal is to model the actors in a way that reflects their activities and their aggregations. In so doing, it would provide a description which would be as complete and accurate as possible of all aspects related to their conduct. The system would then be able to provide dashboards capable of facilitating the task of making accurate decisions [8]. The model adopted should reflect a genuine picture of the system as it uses the UML standard which is the newest in the world of information systems design.

## 2. Hypotheses:

We will start by modeling [4] actors up taking into account the requirements and expectations of each of them, namely:

- The student who wishes to have quality training and be endowed with skills facilitating integration into life.
- The teacher who has the task of producing and transferring knowledge.
- The administration staff whose task is on the one hand to facilitate the work of the teacher serving students, disseminating and sharing information, and on the other hand to meet the needs of clients outside university.

Given this situation, it is a must that the task of the student [1], [2] and that of the teacher and the administrators be correlated. In fact, we are faced with a situation where we are aiming at the satisfaction of the customer / user with a specific university that concept actor / user; also as a business. Indeed, a company's approach to governance is 'profit', while that of a university is about positioning and achieving visibility of the organization. The company seeks a position of performance at its capital and the university aims to achieve quality and a high ranking both nationally and internationally. The company seeks customer satisfaction, whereas the university seeks to satisfy its stakeholders. Customer satisfaction in business

is formalized in terms of costs but satisfaction in university occurs by meeting their needs.

### 3. Context

Today the information system of the Moroccan university produces a large volume of data and information (in [10] [9]). Often, because of this large volume, it becomes difficult to make sense of these data in and out of precise and reliable indicators. To exploit and manage these data, the students, teachers and administration decision makers at the central level do not have standardized data. Most national universities have opted for the 'Apogee' (management tuition). For those teachers and administrators, it offers a dynamic database with open source tools. Thus, our model consists of three-level data warehouses:

- The actor level: it consists of three groups: student, teacher and administrator.

- \*\* Database repository: includes all personal data (rank, specialty, ...)

- Educational level: \*\* Baseline Service instruction (courses, modules ...).

- \*\* Base of regulations: Regulations and Rights of each actor.

- Administrative level: \*\* Administration Database on the situation of all actors.

### 4. Modeling the actors

Previously (in [1] [2]), we showed that applications for the actor level are based on information gathered from the databases (Apogee4, BDens5, BDfon6). The design of a Decision Support Information System [3] requires a special approach to design and modeling complex [4]. We adopted a model to meet specific needs such as factor analysis [5] which has a policy to facilitate understanding and interpretation of a large set of multidimensional data. This analysis shows graphically the similarities between the data and quantifies the degree of correlation between several factors.

The model we get includes all actors involved in the university system. It is follows:

$$\text{Actor} = T_i ; \sum_{i=1}^{i=n} S_i ; \sum_{j=1}^{j=3} C_j ; \sum_{k=1}^{k=n-1} A_k ; \quad (1)$$

With: S: Source activity for all players. C: Category. A: all aggregations. After the consolidation of the formula 1.

Channels	Actors	Level actors	Roles	Activities				Aggregations					
				cycle University		cycle University		cycle University		cycle University			
Study English	Student	Students of the 2nd round	study	Prepare	Administer	Prepare	Registration	Follow	Preparation				
				Learn	Control	Review	Preparation	Preparation	Diploma				
				Follow									
				Prepare	Participate	Enforce	Stage						
				Prepare		Enforce	Control	Participate					
				Prepare									
				Prepare									
				Prepare									
				Prepare									
				Prepare									
				Prepare									
				Prepare									
	Teacher	Research team leader Professor Administer	Teaching	Prepare	Prepare	Preparation	Preparation	Training	Stage				
				Enforce		Training	Training	Control	Training				
				Enforce	Inform		Follow						
				Admin	Training		Preparation						
				Controlled			Stage						
	Administrative	University President Accountant Manager Manager Adviser	Administer Manage Adviser	Administer	Prepare	Training	Budget	Preparation	Participation				
				Enforce	Inform		Registration	Diploma					
				Enforce		SD		Participate	Statistical				
				Admin			Preparation	Training					

Table 1: Role, Activities, Aggregation of Actors

To better understand this approach, we use a graphic to show the equilibrium relationship between each actor and their activities at an undergraduate level [6] and its aggregation, taking into account the multiple observations to develop our model.

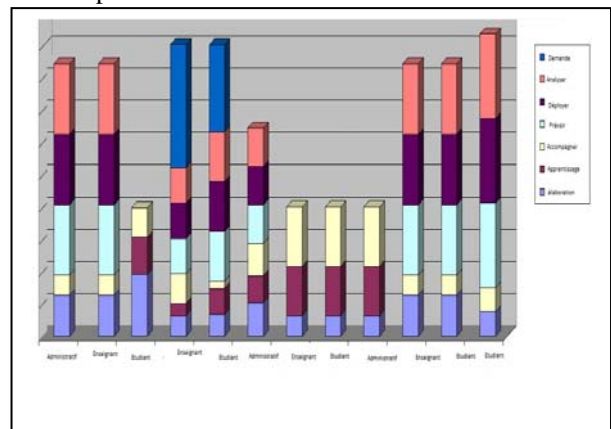


Figure 1: Actors and their activities in relation to their aggregation in a graduate level.

From this graph, it follows that we can help advance each identified actor, by offering additional information that can improve its contribution. Simply correlate all users and information resources during a cycle by a university reunion of all activities of each actor in relation to its aggregation.

## 5- The Dashboard Conception Method

This approach is intended to guide the GIMS system designer performance measurement as a phasing sequence. By proposing to follow the footsteps of the method, the actor wants to encourage the university to think in terms of strategy and objectives before proceeding to the implementation of a performance measurement system. In addition, the indicators must be constructed and selected taking into account not only the goals of the university, but also the activities and aggregations of each player during a university cycle. At all costs, the designer of the process should avoid the dashboard reduction to a predetermined set of synthetic indicators. The process is divided into 10 stages, each covering a particular concern and each marking an identifiable threshold in the system advancement.

1-The steps can be grouped into four main phases:

<p>1 Identification :What is the context?</p>	<ul style="list-style-type: none"> <li>• <b>Step 1: Environment of the University :</b> Analysis of the socioeconomic environment and strategy of the University to define the scope and reach of our study</li> <li>• <b>Step 2: Identification of the University:</b> Structural analysis of the Moroccan University to identify aggregations, activities and actors concerned.</li> </ul>
<p>2 Conception :What is it necessary to make?</p>	<ul style="list-style-type: none"> <li>• <b>Step 3: Definition of objectives:</b> Selection of tactical objectives at the end of each cycle for each team University based on the general strategy</li> <li>• <b>Step 4: Construction of the dashboard:</b> Definition of the dashboard of each presidential team.</li> <li>• <b>Step 5: Selection of indicators:</b> Selection of indicators based on the objectives selected, context and stakeholders</li> <li>• <b>Step 6: Collection of information:</b> Identification of information needed to construct indicators (data available at the base data)</li> <li>• <b>Step 7: The system dashboard:</b> Construction of the system panel, controls the overall consistency</li> </ul>
<p>3 Implementation :How to make him?</p>	<ul style="list-style-type: none"> <li>• <b>Stage 8: the choice of software packages:</b> Elaboration of the sailing(bar) of selection for the choice of the adequate software packages (interest us in produced free)</li> <li>• <b>Stage 9: integration and deployment</b> :Setting-up(Presence) of software packages, deployment at the University groups together (includes) all the establishments</li> </ul>
<p>4 Continuous Improvement, The system always correspond to expectations?</p>	<ul style="list-style-type: none"> <li>• <b>Step 10: Audit</b> :Permanent monitoring system</li> </ul>

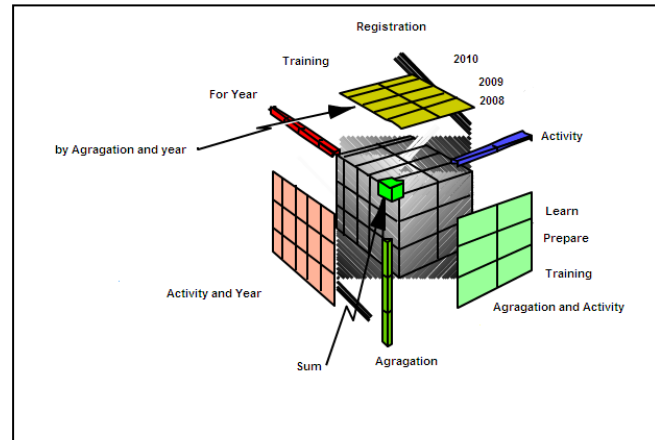
### 2 - Background to our approach

Construction of a multidimensional conceptual model: We use the UML to model the types of actors (teacher, student, and administrator). Indeed, UML has a graphical notation

as a visual based diagram that can facilitate decision-making.

### 3-Data Visualization:

To assist the representation of our model, we visualize for example data about student actors. The figure below shows the activities and aggregations of this actor from the information system according to the academic cycle.



**Figure 2:** Vision of the actor in student activities and Aggregation

### Application of the model by the ‘Pentaho’ open source software.

The purpose of this application is to justify the balance between all activities of all actors and their aggregations at the end of a graduate level.

In this context, we present, as an application, the indicators defined by the decision-makers of the university and planned by technical information system decision-making institution in order to improve the performance of each actor.

#### A - Pre-requisites:

- At the local level (property), the production of dashboard indicators necessitates reliability in both the information system and the expanded local data. They also need to have a service charge of management assistance and management control.
- At the central level (Ministry), it is advisable to define and harmonize the national indicators, to facilitate data exchange and to avoid multiple entries of the same data.

B-The different types of indicators and the different panels, the dashboard of the President are clearly distinct, to clarify the analysis:

- Context indicators, to characterize the institution (statistical key figures);



- Activity indicators, involving data on the management of the institution (examples: credit availability, the rate of commitment expenditures, compliance with job limits and payroll, opening hours of libraries, rate of access to sports facilities, etc..)

- Performance indicators to measure the effectiveness of the institution for purposes defined by the directives and the procedure of the institution (examples: more added value in the curriculum license rate of employability of graduates, proportion of the research team graded A + and A, etc..). Emphasize the link between objectives and indicators. Show the hierarchy and the joints between the various indicators.

C- Expected impact on flying the establishment is presented below in the form of our model with all parameters for the three actors. Decision-makers are to be informed at the University of Ibn Zohr, Agadir, taken as scope.

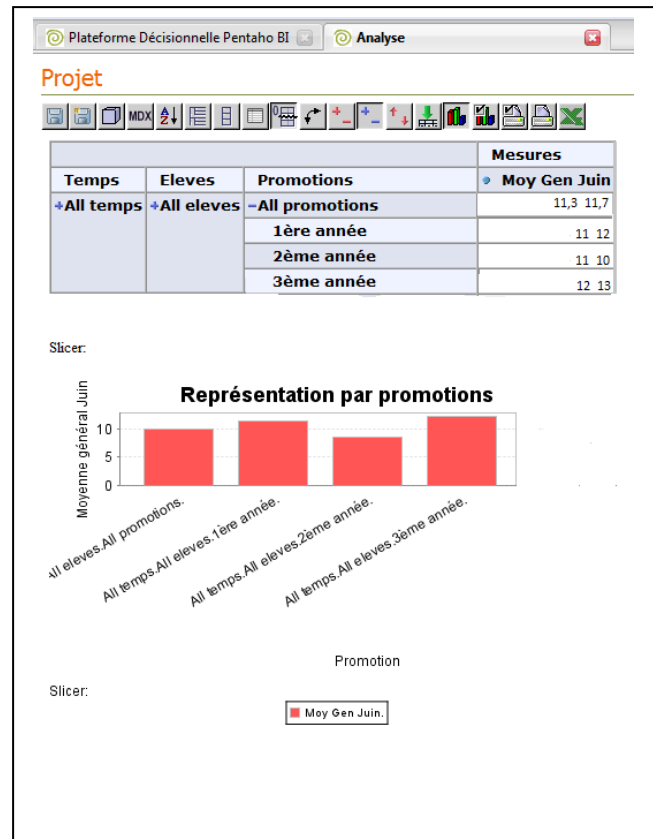


**Figure 3:** model constructed from T = (S, C, A)

This form collects data via a script to feed a mysql database named «ma\_table»; and creates a table of fact called «ma\_table\_fait».

It supplies real-time SQL database and allows the actor to do his own analysis with MDX query in an environment facilitated by the graphical interface of 'Pentaho'. Furthermore, it is possible to modify the MDX query depend criteria defined by decision makers.

The user can then choose which types of graphs are most appropriate to conduct a multidimensional analysis and perform a data mining practice. We are presenting below a screen that shows an analysis made on the basis of available data for the player concerned and student indicator 'success rate at the end of the academic cycle.



**Figure 4:** Evolution success rate in the department of “English Studies” on the basis of the baccalaureate option and academic year

This chart shows the evolution of the success rate in the department of “English Studies” on the basis of the baccalaureate option and academic year

## 6- Conclusion

To implement this application, we went through three major phases. The first is the theoretical part that needs to have a model which is able to meet the academic context known by its complexity (different actors, the wealth of data, non-uniform data ...). This requires a mathematical model defining simple relationships between the actors, their activities and their aggregations. The second phase focuses on gathering data and designing a multidimensional database. The third phase is devoted to the implementation and construction of a scoreboard checking all the proposals made in the theoretical part. The results obtained from available data of the student actor at the university are encouraging. The availability of actual data of the other actors need is a comprehensive decision-making tool for the university.

## 7-References

- [1].Oubedda L, Mir. A, Khalfaoui. M, "Modeling Resources Steering University ", acte SIL' 08, ENSA Marrakech 2008.
- [2]. Oubedda L, Mir. A, Khalfaoui. M, «Human Resource Management of breast Moroccan Universities ", ENCG AGADIR 2009.
- [3].Olivier Bistorin, «Methods and tools for designing business process training systems ", thèse de Doctorat, Université Paul Verlaine-Metz, décembre 2007
- [4].Olivier Bistorin, Thiband Monteiro, Claude Pourcel, " Process modeling of a training system ", Proceedings 1ère Conférence Internationale sur l'Ingénierie des Systèmes de Formation, Carthagène des Andes, Colombie, octobre 2007.
- [5].Peguiro F,David A,Thiery O, " Intelligence in academic settings including the user modeling ", IERA 2003,Nancy.
- [6].Bouaka N. et David A.," Model for the operation of a decision problem: a tool for decision support in a context of economic intelligence ", IERA 2003, Nancy.
- [7].THIERY O., DUCREAU A., BOUAKA N., DAVID A., " Drive an organization: strategic information to the modeling of the user application to the field of HRM ", GREFIGE, 2004.
- [8]. Samia AITOCHE, Abdelghafour KANIT and Kinza MOUSS , PROPOSAL OF A SUPPORT DECISION SUPPORT USING THE METHOD GIMSI, FOR BETTER reactivity in a disturbed environment, International Conference On Industrial Engineering and Manufacturing ICIEM'10, May, 9-10, 2010, Batna, Algeria.
- [9]. Hafid Boutaleb Joutei, A guide to indicators performance, Octobre 2009.
- [10]. Contractual policy in higher education and research and WaveC2009-2012 research Ministry of Higher Education and Research Directorate-General Higher Education.

## 7. Biography of Authors

**OUBEDDA LATIFA**, Student Business Intelligence researcher at the National School of Applied Sciences of Agadir, I have a diploma of Superior Studies Specialized in Business Intelligence at the National School of Applied Sciences of Agadir and Engineering degree in Computer Electronics Electrical Automatic Faculty of Sciences Rabat, areas of research interest and method of application flows and design tools for decision support for the Moroccan university system, and application of the principle of intelligence in the university environment, Database Organization and Operation Research. I hold an engineer position in intelligence at the Faculty of Arts and Humanities AGADIR

**Erraha BRAHIM (PH D)**, Ability Professor in Computer Science at the National School of Applied Sciences of Agadir  
And team member of the Laboratory of Industrial Engineering and Computer Science (LG2I), National School of Applied Sciences of Agadir, University Ibn Zohr.

**KHALFAOUI MOHAMED (PES, D)**, Professor of Higher Education in Applied Mathematics at the School of Technology of Salé, and team member of the Laboratory of Analysis of Systems, Information Processing and Integrated Management , School of Salt technology, University Mohammed V Agdal,

# Outlier Detection: Applications And Techniques

Karanjit Singh and Dr. Shuchita Upadhyaya

HQ Base Workshop Group EME  
Meerut Cantt, UP, India

Department of Computer Science and Applications, Kurukshetra University  
Kurukshetra, Haryana, India

## Abstract

Outliers once upon a time regarded as noisy data in statistics, has turned out to be an important problem which is being researched in diverse fields of research and application domains. Many outlier detection techniques have been developed specific to certain application domains, while some techniques are more generic. Some application domains are being researched in strict confidentiality such as research on crime and terrorist activities. The techniques and results of such techniques are not readily forthcoming. A number of surveys, research and review articles and books cover outlier detection techniques in machine learning and statistical domains individually in great details. In this paper we make an attempt to bring together various outlier detection techniques, in a structured and generic description. With this exercise, we hope to attain a better understanding of the different directions of research on outlier analysis for ourselves as well as for beginners in this research field who could then pick up the links to different areas of applications in details.

**Keywords:** *Outlier Applications, Outliers, Outlier Detection.*

## 1. Introduction

Outlier detection aims to find patterns in data that do not conform to expected behavior. It has extensive use in a wide variety of applications such as military surveillance for enemy activities, intrusion detection in cyber security, fraud detection for credit cards, insurance or health care and fault detection in safety critical systems. Their importance in data is due to the fact that they can translate into actionable information in a wide variety of applications. An anomalous traffic pattern in a computer network could mean that a hacked computer is sending out sensitive data to an unauthorized destination [1]. An abnormal MRI image may indicate presence of malignant tumors [2]. Outliers in credit card transaction data could indicate credit card or identity theft [3] or abnormal readings from a space craft sensor could signify a fault in some component of the space craft [4]. In statistical data study of outliers dates as early as the 19th century [5]. Since then several research communities have developed a variety of outlier detection techniques with many of these

specifically meant for certain applications and others being generic in nature. With this exercise, we hope to get a better understanding of the different directions of research on outlier analysis and think of applying techniques in different areas to our areas of interest of crime detection and counter terrorism, even if they were they were not intended, to begin with.

## 2. Defining Outliers

Outliers are patterns in data that do not conform to a well defined notion of normal behavior. Figure 1 illustrates outliers in a simple 2-dimensional data set. The data has two normal regions,  $N_1$  and  $N_2$ , since most observations lie in these two regions. Points that are sufficiently far away from the regions, e.g., points  $o_1$  and  $o_2$ , and points in region  $O_3$ , are outliers.  $x$   $y$   $N_1$   $N_2$   $o_1$   $o_2$   $O_3$

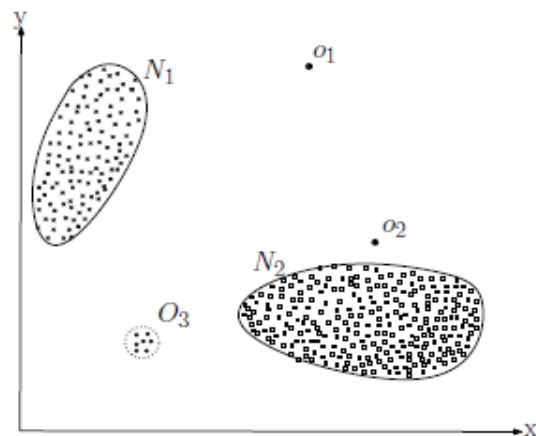


Fig. 1. A simple example of outliers in a 2-dimensional data set.

outliers might be induced in the data for a variety of reasons, such as malicious activity, e.g., credit card fraud, cyber-intrusion, terrorist activity or breakdown of a system,

but the common point of all is that they are interesting to the analyst. The “interestingness” or real life relevance of outliers is a key feature of outlier detection.

Outlier detection is related to, but distinct from noise removal [6] and noise accommodation [7], both of which deal with unwanted noise in the data. Noise can be defined as a phenomenon in data which is not of interest to the analyst, but acts as a hindrance to data analysis. Noise removal is driven by the need to remove the unwanted objects before any data analysis is performed on the data. Noise accommodation refers to immunizing a statistical model estimation against anomalous observations [8].

Another topic related to outlier detection is novelty detection [9,10,11] which aims at detecting previously unobserved (emergent, novel) patterns in the data, e.g., a new topic of discussion in a news group. The distinction between novel patterns and outliers is that the novel patterns are typically incorporated into the normal model after being detected. Another topic related to outlier detection is novelty detection [9,10,11]. The distinction between novel patterns and outliers is that the novel patterns are previously unobserved and get typically incorporated into the normal model after being detected e.g., a new topic of discussion in a news group.

We have discussed above mentioned related problems because their solutions are often used for outlier detection and vice-versa.

### 3. Difficulties in Outlier Detection

Abstractly speaking outliers are patterns that deviate from expected normal behavior, which in its simplest form could be represented by a region and visualize all normal observations to belong to this normal region and consider the rest as outliers This approach looks simple but is highly challenging due to following reasons.

It is very difficult to define the normal behavior or a normal region. The difficulties are as under.

- Encompassing of every possible normal behavior in the region.
- Imprecise boundary between normal and outlier behavior since at times outlier observation lying close to the boundary could actually be normal, and vice-versa.
- Adaptation of malicious adversaries to make the outlier observations appear like normal when outliers result from malicious actions.

- In many domains normal behavior keeps evolving and may not be current to be a representative in the future.
- Differing notion of outliers in different application domains makes it difficult to apply technique developed in one domain to another. For example, in the medical domain a small deviation from normal body temperature might be an outlier, while similar value deviation in the stock market domain might be considered as normal. Even within same domain say crime detection there could be situations where use of foreign make weapons may be considered normal in crimes committed in metro cities but an outlier for murders of commoners in tribal regions.
- Availability of labeled data for training/ validation of models used by outlier detection techniques.
- Noise in the data which tends to be similar to the actual outliers and hence difficult to distinguish and remove.

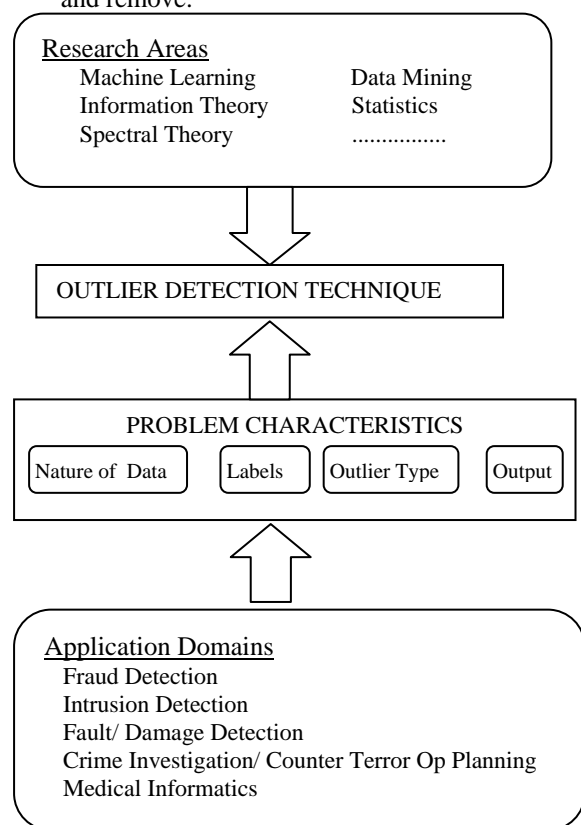


Fig. 2. Key components associated with outlier detection technique.

Due to the above challenges, the outlier detection problem, in its most general form, is not easy to solve. In fact, most of the existing outlier detection techniques solve a specific problem formulation which is induced by various factors

such as nature of the data, availability of labeled data, type of outliers to be detected, etc. Often, these factors are determined by the application domain in which the outliers need to be detected.

Researchers adopt concepts from diverse disciplines such as statistics, machine learning, data mining, information theory, spectral theory, and apply them to specific problem formulations. Figure 2 shows the above mentioned key components associated with any outlier detection technique.

#### 4. Previous Work

A number of surveys, review articles and books especially Hodge and Austin [12], cover outlier detection techniques in machine learning and statistical domains. Numeric and symbolic data approaches [13], neural networks and statistical approaches [9, 10, 11] have been presented by various researchers. Cyber-intrusion detection surveys [14, 15] and research and review books on outlier detection techniques [16,17,18] are excellent sources of literature on the subject.

#### 5. Our Contribution

The above literature on outlier detection set focus on individual applications or on a particular research area. We have attempted to structure and present a broad overview of the detailed research on outlier detection techniques in multifarious research areas and applications also trying to highlight the richness and complexity associated with each application domain. We distinguish simple outliers from complex outliers and define two types of complex outliers, viz., contextual and collective outliers.

#### 6. Aspects Determining the Formulation of Problem

As mentioned earlier, a specific formulation of the problem is determined by several different factors some of which are discussed below. These are also depicted in Figure 2 above. Broadly speaking they are:-

- Nature of Input Data
- Type of Outlier – Point, Contextual, Collective
- Data Labels
- Output of Outlier Detection.

#### 7. Nature of Input Data

This is a key aspect of any outlier detection technique. Input is generally a collection of data instances (also referred as object, record, point, vector, pattern, event, case, sample, observation, entity) [20]. Each data instance can be described using a set of attributes (also referred to as variable, characteristic, feature, field, dimension). The attributes can be of different types such as binary, categorical or continuous. Each data instance might consist of only one attribute (univariate) or multiple attributes (multivariate). In the case of multivariate data instances, all attributes might be of same type or might be a mixture of different data types. The nature of attributes determines the applicability of outlier detection techniques. For example, for statistical techniques different statistical models have to be used for continuous and categorical data. Similarly, for nearest neighbor based techniques, the nature of attributes would determine the distance measure to be used. Often, instead of the actual data, the pair-wise distance between instances might be provided in the form of a distance (or similarity) matrix. In such cases, techniques that require original data instances are not applicable, e.g., many statistical and classification based techniques. In case these statistical methods are applied to OLAP cubes for data mining then the distance between dimensional data can be found out by applying score functions.

Input data can also be categorized based on the relationship present among data instances [20]. Most of the existing outlier detection techniques deal with record data (or point data), in which no relationship is assumed among the data instances. In case these statistical methods are applied to OLAP cubes for data mining then the distance between dimensional data can be found out by applying some sort of score functions and then determining the outliers.

In general, data instances can be related to each other. Some examples are sequence data, spatial data, and graph data. In sequence data, the data instances are linearly ordered, e.g., time-series data, genome sequences, protein sequences. In spatial data, each data instance is related to its neighboring instances, e.g., vehicular traffic data, ecological data. When the spatial data has a temporal (sequential) component it is referred to as spatio-temporal data, e.g., climate data. In graph data, data instances are represented as vertices in a graph and are connected to other vertices with edges. Later we will discuss situations where such relationship among data instances becomes relevant for outlier detection.

## 8. Types of Outliers

An important aspect of an outlier detection technique is the nature of the desired outlier. Outliers can be classified into following three categories:

- Point Outliers
- Contextual Outliers
- Collective Outliers.

## 9. Point Outliers

If an individual data instance can be considered as anomalous with respect to the rest of data, then the instance is termed as a point outlier. This is the simplest type of outlier and is the focus of majority of research on outlier detection. For example, in Figure 1, points o1 and o2 as well as points in region O3 lie outside the boundary of the normal regions, and hence are point outliers since they are different from normal data points. As a real life example, if we consider credit card fraud detection with data set corresponding to an individual's credit card transactions assuming data definition by only one feature: amount spent. A transaction for which the amount spent is very high compared to the normal range of expenditure for that person will be a point outlier.

## 10. Contextual Outliers

If a data instance is anomalous in a specific con-text (but not otherwise), then it is termed as a contextual outlier (also referred to as conditional outlier [21]). The notion of a context is induced by the structure in the data set and has to be specified as a part of the problem formulation. Each data instance is defined using two sets of attributes:

- **Contextual attributes.** The contextual attributes are used to determine the context (or neighborhood) for that instance. For example, in spatial data sets, the longitude and latitude of a location are the contextual attributes. In time-series data, time is a contextual attribute which determines the position of an instance on the entire sequence.
- **Behavioral attributes.** The behavioral attributes define the non-contextual characteristics of an instance. For example, in a spatial data set describing the average rainfall of the entire world, the amount of rainfall at any location is a behavioral attribute.

The anomalous behavior is determined using the values for the behavioral attributes within a specific context. A data instance might be a contextual outlier in a given context, but an identical data instance (in terms of behavioral attributes) could be considered normal in a different context. This property is key in identifying contextual and behavioral attributes for a contextual outlier detection technique.

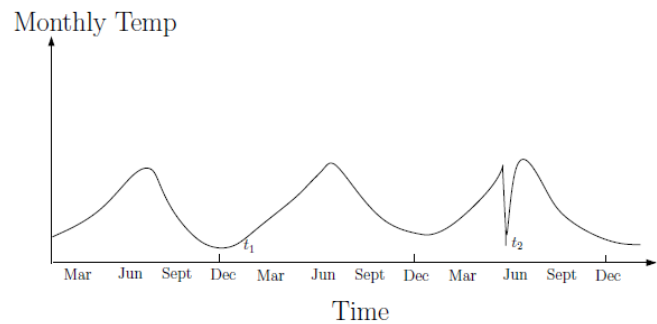


Fig. 3. Contextual outlier t2 in a temperature time series. Temperature at time t1 is same as that at time t2 but occurs in a different context and hence is not considered as an outlier.

Contextual outliers have been most commonly explored in time-series data [22] and spatial data [23]. Figure 3 shows one such example for a temperature time series which shows the monthly temperature of an area over last few years. A temperature of 35F might be normal during the winter (at time  $t_1$ ) at that place, but the same value during summer (at time  $t_2$ ) would be an outlier. A six ft tall adult may be a normal person but if viewed in *context of age* a *six feet* tall *kid* will definitely be an outlier.

A similar example can be found in the credit card fraud detection with contextual as *time* of purchase. Suppose an individual usually has a weekly shopping bill of \$100 except during the Christmas week, when it reaches \$1000. A new purchase of \$1000 in a week in July will be considered a contextual outlier, since it does not conform to the normal behavior of the individual in the context of time (even though the same amount spent during Christmas week will be considered normal).

The choice of applying a contextual outlier detection technique is determined by the meaningfulness of the contextual outliers in the target application domain. Applying a contextual outlier detection technique makes sense if contextual attributes are readily available and therefore defining a context is straightforward. But it becomes difficult to apply such techniques if defining a context is not easy.

## 11. Collective Outliers

If a collection of related data instances is anomalous with respect to the entire data set, it is termed as a collective outlier. The individual data instances in a collective outlier may not be outliers by themselves, but their occurrence together as a collection is anomalous. Figure 4 illustrates an example which shows a human electrocardiogram output [24]. The highlighted region denotes an outlier because the same low value exists for an abnormally long time (corresponding to an Atrial Premature Contraction). It may be noted that low value by itself is not an outlier but its successive occurrence for long time is an outlier.

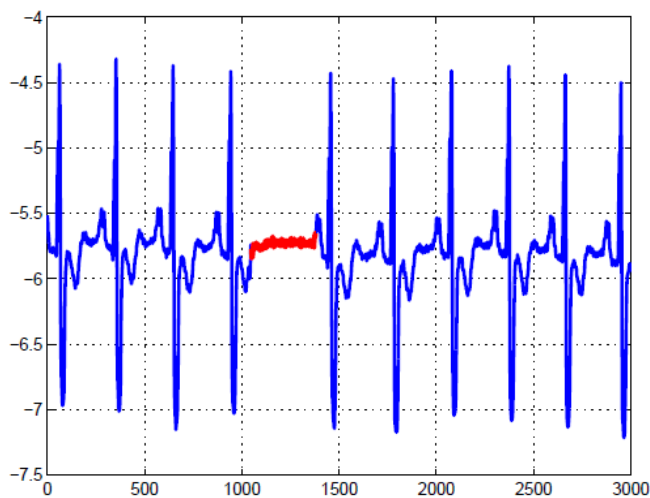


Fig. 4. Collective outlier in an human ECG output corresponding to an Atrial Premature Contraction.

As an another illustrative example, consider a sequence of actions occurring in a computer as shown below:

.....http-web, buffer-overflow, http-web, http-web, smtp-mail, ftp, http-web, ssh, smtp-mail, http-web, ssh, buffer-overflow, ftp, http-web, ftp, smtp-mail, http-web.....

The highlighted sequence of events (buffer-overflow, ssh, ftp) correspond to a typical web based attack by a remote machine followed by copying of data from the host computer to remote destination via ftp. It should be noted that this collection of events is an outlier but the individual events are not outliers when they occur in other locations in the sequence.

Collective outliers have been explored for sequence data [25,26], graph data [27], and spatial data [28]. It should be noted that while point outliers can occur in any data set, collective outliers can occur only in data sets in which data instances are related. In contrast, occurrence of contextual outliers depends on the availability of context attributes in

the data. A point outlier or a collective outlier can also be a contextual outlier if analyzed with respect to a context. Thus a point outlier detection problem or collective outlier detection problem can be transformed to a contextual outlier detection problem by incorporating the context information

## 12. Data Labels

The labels associated with a data instance denote if that instance is normal or anomalous. It should be noted that obtaining labeled data which is accurate as well as representative of all types of behaviors, is often prohibitively expensive. Labeling is often done manually by a human expert and hence requires substantial effort to obtain the labeled training data set. Typically, getting a labeled set of anomalous data instances which cover all possible type of anomalous behavior is more difficult than getting labels for normal behavior. Moreover, the outlier behavior is often dynamic in nature, e.g., new types of outliers might arise, for which there is no labeled training data. In certain cases, such as air traffic safety, outlier instances would translate to catastrophic events, and hence will be very rare. Based on the extent to which the labels are available, outlier detection techniques can operate in one of the following three modes:

- **Supervised outlier detection:** Techniques trained in supervised mode assume the availability of a training data set which has labeled instances for normal as well as outlier class. Typical approach in such cases is to build a predictive model for normal vs. outlier classes. Any unseen data instance is compared against the model to determine which class it belongs to. There are two major issues that arise in supervised outlier detection. First, the anomalous instances are few, as compared to the normal instances in the training data. Second, obtaining accurate and representative labels, especially for the outlier class is usually challenging. A number of techniques have been proposed [ 29, 30, 31] that inject artificial outliers in a normal data set to obtain a labeled training data set. Other than these two issues, the supervised outlier detection problem is similar to building predictive models. Hence we will not address this category of techniques in this survey.
- **Semi-Supervised outlier detection:** Techniques that operate in a semi-supervised mode, assume that the training data has labeled instances for only the normal class. Since they do not require labels for the outlier class, they are more widely

applicable than supervised techniques. For example, in space craft fault detection [32], an outlier scenario would signify an accident, which is not easy to model. The typical approach used in such techniques is to build a model for the class corresponding to normal behavior, and use the model to identify outliers in the test data. A limited set of outlier detection techniques exist that assume availability of only the outlier instances for training [25, 33, 34]. Such techniques are not commonly used, primarily because it is difficult to obtain a training data set which covers every possible anomalous behavior that can occur in the data.

- **Unsupervised outlier detection:** Techniques that operate in unsupervised mode do not require training data, and thus are most widely applicable. The techniques in this category make the implicit assumption that normal instances are far more frequent than outliers in the test data. If this assumption is not true then such techniques suffer from high false alarm rate.

Many semi-supervised techniques can be adapted to operate in an unsupervised mode by using a sample of the unlabeled data set as training data. Such adaptation assumes that the test data contains very few outliers and the model learnt during training is robust to these few outliers.

### 13. Output of Outlier Detection

An important aspect for any outlier detection technique is the manner in which the outliers are reported. Typically, the outputs produced by outlier detection techniques are one of the following two types:

- **Scores:** Scoring techniques assign an outlier score to each instance in the test data depending on the degree to which that instance is considered an outlier. Thus the output of such techniques is a ranked list of outliers. An analyst may choose to either analyze top few outliers or use a cut-off threshold to select the outliers.
- **Labels:** Techniques in this category assign a label (normal or anomalous) to each test instance.

Scoring based outlier detection techniques allow the analyst to use a domain-specific threshold to select the most relevant outliers. Techniques that provide binary labels to the test instances do not directly allow the analysts to make such a choice, though this can be

controlled indirectly through parameter choices within each technique.

### 14. Applications of Outlier Detection

We shall highlight several applications of outlier detection. For each application we shall discuss following aspects:

- The notion of outlier.
- Nature of the data.
- Challenges associated with detecting outliers.
- Existing outlier detection techniques.

### 15. Intrusion Detection

Intrusion detection refers to detection of malicious activity (break-ins, penetrations, and other forms of computer abuse) in a computer related system [35] interesting from a computer security perspective. Being different from normal system behavior, intrusion detection is a perfect candidate for applying outlier detection techniques. The key challenges for outlier detection are :-

- **Huge Data Volume:** This calls for computationally efficient techniques.
- **Streaming Data:** This requires on-line analysis.
- **False alarm rate:** Smallest percentage of false alarms among millions of data objects can make be overwhelming for an analyst.
- **Labeled data not usually available for Intrusions:** This gives preference to semi-supervised and unsupervised outlier detection techniques.

Intrusion detection systems have been classified into host based and network based intrusion detection systems [ 36].



The major differences being tabled as under:-

Table 1: Differences in Nature of Host Based and Network Intrusion Systems

<i>Aspect</i>	<i>Host Based</i>	<i>Network Based</i>
Outliers In	OS Calls	Network Data.
Translates to	Malicious Code Unusual Behaviour Policy Violations	Denial of Network Services
Nature of Data analysis	Sequential	Point, Sequential, Collective
Granularity/ Profiling	User / Program	Packet Level/ NetFlows

The examples of outlier detection techniques for Intrusion Detection are tabled below:-

Table 2: Some outlier detection techniques used in Host Based and Network Intrusion Systems

<i>Technique Used</i>	<i>References</i>
<b>Host Based Intrusion Detection Systems</b>	
Statistical Profiling Using Histograms	[37- 45]
Mixture of Models	[46]
Neural Networks	[47]
Support Vector Machines	[2]
Rule Based Systems	[48 - 50]
<b>Network Based Intrusion Detection Systems</b>	
Statistical Profiling using Histograms	[51 - 54]
Parametric Statistical Modeling	[55]
Non-parametric Statistical Modeling	[56]
Bayesian Networks	[57 - 60]
Support Vector Machines	[46]
Rule Based Systems	[61]
Neural Networks	[46, 62 - 68]

## 16. Fraud Detection

Fraud refers to criminal activities occurring in commercial organizations such as banks, credit card companies,

insurance agencies, cell phone companies, stock market, etc. Malicious users could be actual customers of the organization or resorting to identity theft (posing as customers). The detection activity aims at detection of unauthorized consumption of resources provided by the organization to prevent economic losses.

A general approach to outlier detection here would involve maintaining a usage profile for each customer and monitor the profiles to detect any deviations termed as activity monitoring [73]. Some specific applications of fraud detection are discussed below.

**Credit Card Fraud Detection:** Outlier detection techniques are applied to detect :-

- **Fraudulent Applications for Credit Card:** This is similar to detecting insurance fraud [69]
- **Fraudulent Usage of Credit Card:** Associated with credit card thefts.

The data records are defined over several dimensions such as the user ID, spent amount, time between consecutive card usage, etc. The frauds are typically reflected in transactional records (point outliers) and correspond to high payments, high rate of purchase, purchase of items never purchased by the user before, etc. Availability of labeled records is no problem since credit companies have complete data available. Moreover, the data falls into distinct profiles based on the credit card user. Hence profiling and clustering based techniques are typically used in this domain.

Online detection of fraud as soon as fraudulent transaction occurs is a challenge in detecting unauthorized credit card usage. This problem is addressed in two different ways.

Table 3: Approaches in detecting fraudulent transactions.

<i>Approach</i>	<i>By-Owner</i>	<i>By-Operation</i>
Context	User	Geographic Location
Cost	Expensive; querying a central data repository with every transaction.	

Some outlier detection techniques used in fraud detection are listed in Table IV.

Table IV: Some outlier detection techniques used in fraud detection

<i>Technique Used</i>	<i>References</i>
Neural Networks	[3, 69 - 71]
Rule-based Systems	[70]
Clustering	[72]

### 17. Mobile Phone Fraud Detection.

In this activity monitoring problem the calling behavior of each account is scanned to issue an alarm when an account appears to have been misused.

Calling activity is usually represented with call records. Each call record is a vector of continuous (e.g., Call-Duration) and discrete (e.g., Calling-City) features. However, there is no inherent primitive representation in this domain. Calls are aggregated by time, for example into call-hours or call-days or user or area depending on the granularity desired. The outliers correspond to high volume of calls or calls made to unlikely destinations.

Some techniques applied to cell phone fraud detection are listed in Table V.

Table V: Examples of different outlier detection techniques used for cell phone fraud detection.

<i>Technique Used</i>	<i>References</i>
Statistical Profiling using Histograms	[73, 74]
Parametric Statistical Modelling	[75, 76]
Neural Networks	[77, 78]
Rule based Systems	[78,79]

### 18. Insurance Claim Fraud Detection

An important problem in the property-casualty insurance industry is claims fraud, e.g. automobile insurance fraud. Individuals and conspiratorial rings of claimants and providers manipulate the claim processing system for unauthorized and illegal claims.

The data in this domain for fraud detection comes from the documents submitted by the claimants. The

techniques extract different features (both categorical as well as continuous) from these documents. Typically, claim adjusters and investigators assess these claims for frauds. These manually investigated cases are used as labeled instances by supervised and semi-supervised techniques for insurance fraud detection.

Insurance claim fraud detection is quite often handled as a generic activity monitoring problem [73]. Neural network based techniques have also been applied to identify anomalous insurance claims [80, 81].

### 19. Insider Trading Detection

Insider trading is a phenomenon found in stock markets, where people make illegal profits by acting on (or leaking) inside information before the information is made public.

The inside information can be of different forms [82] generally referring to any information which would affect the stock prices in a particular industry. It could be knowledge about a pending merger/acquisition, a terrorist attack affecting a particular industry, a pending legislation affecting a particular industry.

Fraud has to be detected in an online manner and as early as possible, to prevent people/organizations from making illegal profits. The available data comes from heterogeneous sources such as option trading data, stock trading data, news. The data has temporal associations since the data is collected continuously. The temporal and streaming nature has also been exploited in certain techniques [75].

Some outlier detection techniques used in this domain are listed in Table VI.

Table VI: Examples of different outlier detection techniques used for insider trading detection.

<i>Technique Used</i>	<i>References</i>
Statistical profiling using Histograms	[75, 82]
Information Theoretic	[83]

### 20. Medical and Public Health Outlier Detection

The data typically consists of patient records which may have several different types of features such as patient age, blood group, weight. The data might also

have temporal as well as spatial aspect to it. The data can have outliers due to several reasons such as abnormal patient condition or instrumentation errors or recording errors. Most of the current outlier detection techniques in this domain aim at detecting anomalous records (point outliers). Typically the labeled data belongs to the healthy patients, hence most of the techniques adopt semi-supervised approach. Another form of data handled by outlier detection techniques in this domain is time series data, such as Electrocardiograms (ECG) and Electroencephalograms (EEG). Collective outlier detection techniques have been applied to detect outliers in such data [91]. Several techniques have also focussed on detecting disease outbreaks in a specific area [90]. Thus the outlier detection is a very critical problem in this domain and requires high degree of accuracy.

The most challenging aspect of the outlier detection problem in this domain is that the cost of classifying an outlier as normal can be very high.

Some outlier detection techniques used in this domain are listed in Table VII.

Table VII: Examples of different outlier detection techniques used in medical and public health domain.

<i>Technique Used</i>	<i>References</i>
Parametric Statistical Modelling	[84 - 88]
Neural Networks	[89]
Bayesian Networks	[90]
Rule-based Systems	[75]
Nearest Neighbor based techniques	[91]

## 21. Industrial Damage Detection

Industrial units suffer damage due to continuous usage and the normal wear and tear. Such damages need to be detected early to prevent further escalation and losses. The data in this domain is usually sensor data recorded using different sensors and collected for analysis.

Outlier detection in this domain is classified into two fields as tabulated below.

Table VIII: Characteristics of Fault Detection in Mechanical Units and Structural Damage Domain.

<i>Aspect</i>	<i>System Health Management</i>	<i>By-Operation</i>
Defects Dealt pertaining to	Mech components such as motors, engines, turbines, oil flow in pipelines etc.	Structures,
Cause of Defects	Wear and Tear or other unforeseen circumstances.	Cracks in beams, strains in airframes . Unforeseen data.
Data Aspect	Temporal	Temporal
Analysis	Time Series	Time series with special correlations
Types of Outliers	Contextual or Collective outliers	Novelty detection or change point detections
Normal data	Readily Available	Is learnt and typically static over time.
Supervision	Semi-supervised	Semi-supervised
Literature	[94, 95]	[108, 111, 112, 115]
Techniques	Table IX.	Table X.

Table IX: Examples of outlier detection techniques used for fault detection in mechanical units.

<i>Technique Used</i>	<i>References</i>
Parametric Statistical Modelling	[92, 93, 94, 95]
Non-Parametric Statistical Modelling	[96]
Neural Networks	[97, 89, 98-105]
Spectral	[4, 106]
Rule Based Systems	[107]

Table X: Examples of outlier detection techniques used for structural damage detection.

<i>Technique Used</i>	<i>References</i>
Statistical profiling using Histograms	[108, 109, 110]
Parametric Statistical Modelling	[111]
Mixture of Models	[112, 113, 114]
Neural Networks	[115 to 122]

## 22. Image Processing

Outlier detection here aims to detect changes in an image over time (motion detection) or in regions which appear abnormal on the static image. This domain includes satellite imagery, digit recognition, spectroscopy, mammographic image, and video surveillance. The outliers are caused by motion or insertion of foreign object or instrumentation errors. The data has spatial as well as temporal characteristics. Each data point has a few continuous attributes such as color, lightness, texture, etc. The interesting outliers are either anomalous points or regions in the images (point and contextual outliers).

One of the key challenges in this domain is the large size of the input. The challenge is greater when dealing with video data and, online detection techniques are required.

Some references on various applications are tabulated below:-

Table XI: Examples of outlier detection techniques used in image processing domain.

<i>Application Domain</i>	<i>References</i>
Satellite Imagery	[123, 124, 125, 126, 127]
Digit Recognition	[128]
Mammographic Image Analysis	[129, 130]
Spectroscopy	[131, 132, 133, 134]
Video Surveillance	[135, 136, 137].

Some outlier detection techniques used in this domain are listed in Table XII.

Table XII: Examples of outlier detection techniques used in image processing domain.

<i>Technique Used</i>	<i>References</i>
Mixture of Models	[124, 129, 130]
Regression	[126, 131]
Bayesian Networks	[135]
Support Vector Machines	[132, 138]
Neural Networks	[123, 125, 128, 133, 136]
Clustering	[134]
Nearest Neighbour Techniques	[124, 137 ]

## 23. Outlier Detection in Text Data

Outlier detection techniques in this domain primarily detect novel topics or events or news stories in a collection of documents or news articles. The outliers are caused due to a new interesting event or an anomalous topic. The data in this domain is typically high dimensional and very sparse. The data also has a temporal aspect since the documents are collected over time.

A challenge for outlier detection techniques in this domain is to handle the large variations in documents belonging to one category or topic. Some outlier detection techniques used in this domain are listed in Table XIII.

Table XIII: Examples of techniques used for outlier topic detection in text data.

<i>Technique Used</i>	<i>References</i>
Statistical Profiling using Histograms	[73]
Mixture of Models	[139]
Neural Networks	[140]
Support Vector Machines	[141]
Clustering Based	[142, 143, 144]

## 24. Sensor Networks

Sensor networks have lately become an important topic of research from data analysis perspective, since the data collected from various wireless sensors has several unique characteristics. Outliers in such data collected can either imply one or more faulty sensors (sensor fault detection applications), or the sensors are detecting events (intrusion detection applications) that are interesting for analysts.

A single sensor network might comprise a mix of sensors that collecting different types of data, such as binary, discrete, continuous, audio, video, etc. The data is generated in a streaming mode and the collected data often contains noise and missing values due to limitations imposed by deployment environment and communication channel.

This poses a set of unique challenges. The streaming data calls for outlier detection techniques to operate in an online approach. The severe resource

constraints call for light-weight detection techniques. The data collected in a distributed fashion calls for a distributed data mining approach to analyze the data [145]. Lastly the presence of noise in sensor data makes outlier detection more challenging, since it has to now distinguish between interesting outliers and the unwanted values (noise/missing values).

Table XIV lists some outlier detection techniques used in this domain.

Table XIV: Some outlier detection techniques used for outlier detection in sensor networks.

<i>Technique Used</i>	<i>References</i>
Bayesian Networks	[146]
Rule-based Systems	[147]
Parametric Statistical Modelling	[148, 149]
Nearest Neighbor Based Techniques	[150, 151, 152]
Spectral Techniques	[145]

## 25. Other Domains

Some other domains where outlier detection has also been applied are as tabulated below.

Table XV: Examples of outlier detection techniques used in other application domains.

<i>Technique Used</i>	<i>References</i>
Speech Recognition	[153, 154]
Novelty Detection in Robot Behavior	[155, 156, 157, 158, 159]
Traffic Monitoring	[160]
Click Through Protection	[161]
Detecting Faults in Web Applications	[162, 163]
Detecting Outliers in Biological Data	[55, 164, 165, 166, 167, 176]
Detecting Outliers in Census Data	[168]
Detecting Associations among Criminal Activities	[169]

Detecting Outliers in Customer Relationship Management (CRM) Data	[170]
Detecting Outliers in Astronomical Data	[171, 172, 173]
Detecting Ecosystem Disturbances	[23,174, 175]

## 26. Conclusion

In this paper we have brought together various outlier detection techniques, in a structured and generic description. With this exercise, we have attained a better understanding of the different directions of research on outlier analysis for ourselves as well as for beginners in this research field who can pick up the links to different areas of applications in details.

## Acknowledgments

Karanjit Singh and Dr Shuchita Upadhyaya thanks the staff of CS and IT department of Kurukshetra University for their wholehearted support in referencing the study material. The authors sincerely thank the computer centre and library staff for their unstinted support as well.

## References

- [1] Kumar, V. 2005. Parallel and Distributed Computing for Cybersecurity. Distributed Systems Online, IEEE 6, 10.
- [2] Spence, C., Parra, L., and Sajda, P. 2001. Detection, synthesis and compression in mammographic image analysis with a hierarchical image probability model. In Proceedings of the IEEE Workshop on Mathematical Methods in Biomedical Image Analysis. IEEE Computer Society, Washington, DC, USA, 3.
- [3] Aleskerov, E., Freisleben, B., and Rao, B. 1997. Cardwatch: A neural network based database mining system for credit card fraud detection. In Proceedings of IEEE Computational Intelligence for Financial Engineering. 220-226.
- [4] Fujimaki, R., Yairi, T., and Machida, K. 2005. An approach to spacecraft outlier detection problem using kernel feature space. In Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. ACM Press, New York, NY, USA, 401-410
- [5] Edgeworth, F. Y. 1887. On discordant observations. Philosophical Magazine 23, 5, 364 -375.
- [6] Teng, H., Chen, K., and Lu, S. 1990. Adaptive real-time outlier detection using inductively generated sequential patterns. In Proceedings of IEEE Computer Society Symposium on Re-search in Security and Privacy. IEEE Computer Society Press, 278-284.

- [7] Rousseeuw, P. J. and Leroy, A. M. 1987. Robust regression and outlier detection. John Wiley & Sons, Inc., New York, NY, USA.
- [8] Huber, P. 1974. Robust Statistics. Wiley, New York.
- [9] Markou, M. and Singh, S. 2003a. Novelty detection: a review-part 1: statistical approaches. *Signal Processing* 83, 12, 2481-2497.
- [10] Markou, M. and Singh, S. 2003b. Novelty detection: a review-part 2: neural network based approaches. *Signal Processing* 83, 12, 2499-2521.
- [11] Saunders, R. and Gero, J. 2000. The importance of being emergent. In *Proceedings of Artificial Intelligence in Design*.
- [12] Hodge, V. and Austin, J. 2004. A survey of outlier detection methodologies. *Artificial Intelligence Review* 22, 2, 85-126
- [13] Agyemang, M., Barker, K., and Alhaji, R. 2006. A comprehensive survey of numeric and symbolic outlier mining techniques. *Intelligent Data Analysis* 10, 6, 521-538.
- [14] Patcha, A. and Park, J.-M. 2007. An overview of outlier detection techniques: Existing solutions and latest technological trends. *Comput. Networks* 51, 12, 3448-3470
- [15] Snyder, D. 2001. Online intrusion detection using sequences of system calls. M.S. thesis, Department of Computer Science, Florida State University.
- [16] Rousseeuw, P. J. and Leroy, A. M. 1987. Robust regression and outlier detection. John Wiley & Sons, Inc., New York, NY, USA.
- [17] Barnett, V. and Lewis, T. 1994. Outliers in statistical data. John Wiley and sons.
- [18] Bakar, Z., Mohamad, R., Ahmad, A., and Deris, M. 2006. A comparative study for outlier detection techniques in data mining. *Cybernetics and Intelligent Systems*, 2006 IEEE Conference, 1- 6
- [19] Varun Chandola, Arindam Banerjee, Vipin Kumar 2009, Outlier Detection, University of Miniesota
- [20] Tan, P.-N., Steinbach, M., and Kumar, V. 2005. Introduction to Data Mining. Addison-Wesley.
- [21] Song, X., Wu, M., Jermaine, C., and Ranka, S. 2007. Conditional outlier detection. *IEEE Transactions on Knowledge and Data Engineering* 19, 5, 631-645.
- [22] Weigend, A. S., Mangeas, M., and Srivastava, A. N. 1995. Nonlinear gated experts for time-series - discovering regimes and avoiding overfitting. *International Journal of Neural Systems* 6, 4, 373-399.
- [23] Kou, Y., Lu, C.-T., and Chen, D. 2006. Spatial weighted outlier detection. In *Proceedings of SIAM Conference on Data Mining*.
- [24] Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. 2000. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* 101, 23, e215 - e220. *Circulation Electronic* Pages: <http://circ.ahajournals.org/cgi/content/full/101/23/e215>
- [25] Forrest, S., Warrender, C., and Pearlmutter, B. 1999. Detecting intrusions using system calls: Alternate data models. In *Proceedings of the 1999 IEEE ISRSP*. IEEE Computer Society, Washington, DC, USA, 133 - 145.
- [26] Sun, P., Chawla, S., and Arunasalam, B. 2006. Mining for outliers in sequential databases. In *SIAM International Conference on Data Mining*.
- [27] Noble, C. C. and Cook, D. J. 2003. Graph-based outlier detection. In *Proceedings of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM Press, 631 - 636.
- [28] Sekar, R., Bendre, M., Dhurjati, D., and Bollineni, P. 2001. A fast automaton-based method for detecting anomalous program behaviors. In *Proceedings of the IEEE Symposium on Security and Privacy*. IEEE Computer Society, 144.
- [29] Theiler, J. and Cai, D. M. 2003. Resampling approach for outlier detection in multispectral images. In *Proceedings of SPIE 5093*, 230-240, Ed.
- [30] Abe, N., Zadrozny, B., and Langford, J. 2006. Outlier detection by active learning. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM Press, New York, NY, USA, 504 - 509.
- [31] Steinwart, I., Hush, D., and Scovel, C. 2005. A classification framework for outlier detection. *Journal of Machine Learning Research* 6, 211 - 232
- [32] Fujimaki, R., Yairi, T., and Machida, K. 2005. An approach to spacecraft outlier detection problem using kernel feature space. In *Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM Press, New York, NY, USA, 401 - 410
- [33] Dasgupta, D. and Nino, F. 2000. A comparison of negative and positive selection algorithms in novel pattern detection. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*. Vol. 1. Nashville, TN, 125 - 130.
- [34] Dasgupta, D. and Majumdar, N. 2002. Outlier detection in multidimensional data using negative selection algorithm. In *Proceedings of the IEEE Conference on Evolutionary Computation*. Hawaii, 1039 - 1044.
- [35] Phoha, V. V. 2002. The Springer Internet Security Dictionary. Springer-Verlag.
- [36] Denning, D. E. 1987. An intrusion detection model. *IEEE Transactions of Software Engineering* 13, 2, 222 - 232.
- [37] Forrest, S., D'haeseleer, P., and Helman, P. 1996. An immunological approach to change detection: Algorithms, analysis and implications. In *Proceedings of the 1996 IEEE Symposium on Security and Privacy*. IEEE Computer Society, 110.
- [38] Forrest, S., Esponda, F., and Helman, P. 2004. A formal framework for positive and negative detection schemes. In *IEEE Transactions on Systems, Man and Cybernetics, Part B*. IEEE, 357 - 373.
- [39] Forrest, S., Hofmeyr, S. A., Somayaji, A., and Longstaff, T. A. 1996. A sense of self for Unix processes. In *Proceedings of the ISRSP96*. 120 - 128.
- [40] Forrest, S., Perelson, A. S., Allen, L., and Cherkuri, R. 1994. Self nonself discrimination in a computer. In *Proceedings of the 1994 IEEE Symposium on Security and*

- Privacy. IEEE Computer Society, Washington, DC, USA, 202.
- [41] Forrest, S., Warrender, C., and Pearlmuter, B. 1999. Detecting intrusions using system calls: Alternate data models. In Proceedings of the 1999 IEEE ISRSP. IEEE Computer Society, Washington, DC, USA, 133 - 145.
- [42] Hofmeyr, S. A., Forrest, S., and Somayaji, A. 1998. Intrusion detection using sequences of system calls. *Journal of Computer Security* 6, 3, 151 - 180.
- [43] Jagadish, H. V., Koudas, N., and Muthukrishnan, S. 1999. Mining deviants in a time series database. In Proceedings of the 25th International Conference on Very Large Data Bases. Morgan Kaufmann Publishers Inc., 102 - 113.
- [44] Cabrera, J. B. D., Lewis, L., and Mehra, R. K. 2001. Detection and classification of intrusions and faults using sequences of system calls. *SIGMOD Records* 30, 4, 25 - 34.
- [45] Gonzalez, F. A. and Dasgupta, D. 2003. Outlier detection using real-valued negative selection. *Genetic Programming and Evolvable Machines* 4, 4, 383- 403.
- [46] Eskin, E. 2000. Outlier detection over noisy data using learned probability distributions. In Proceedings of the Seventeenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., 255 - 262.
- [47] Ghosh, A. K., Wanken, J., and Charron, F. 1998. Detecting anomalous and unknown intrusions against programs. In Proceedings of the 14th Annual Computer Security Applications Conference. IEEE Computer Society, 259
- [48] Lee, W. and Stolfo, S. 1998. Data mining approaches for intrusion detection. In Proceedings of the 7th USENIX Security Symposium. San Antonio, TX.
- [49] Lee, W., Stolfo, S., and Chan, P. 1997. Learning patterns from Unix process execution traces for intrusion detection. In Proceedings of the AAAI 97 workshop on AI methods in Fraud and risk management.
- [50] Lee, W., Stolfo, S. J., and Mok, K. W. 2000. Adaptive intrusion detection: A data mining approach. *Artificial Intelligence Review* 14, 6, 533 - 567.
- [51] Anderson, Lunt, Javitz, Tamaru, A., and Valdes, A. 1995. Detecting unusual program behavior using the statistical components of NIDES. Tech. Rep. SRI - CSL - 95 - 06, Computer Science Laboratory, SRI International.
- [52] Anderson, D., Frivold, T., Tamaru, A., and Valdes, A. 1994. Next-generation intrusion detection expert system (NIDES), software users manual, beta-update release. Tech. Rep. SRI CSL - 95 - 07, Computer Science Laboratory, SRI International.
- [53] Porras, P. A. and Neumann, P. G. 1997. EMERALD: Event monitoring enabling responses to anomalous live disturbances. In Proceedings of 20th NIST-NCSC National Information Systems
- [54] Yamanishi, K. and Ichi Takeuchi, J. 2001. Discovering outlier filtering rules from unlabeled data: combining a supervised learner with an unsupervised learner. In Proceedings of the 7th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM Press, 389 - 394.
- [55] Gwadera, R., Atallah, M. J., and Szpankowski, W. 2005b. Reliable detection of episodes in event sequences. *Knowledge and Information Systems* 7, 4, 415 - 437.
- [56] Chow, C. and Yeung, D.-Y. 2002. Parzen-window network intrusion detectors. In Proceedings of the 16th International Conference on Pattern Recognition. Vol. 4. IEEE Computer Society, Washington, DC, USA, 40385.
- [57] Siaterlis, C. and Maglaris, B. 2004. Towards multisensor data fusion for DoS detection. In Proceedings of the 2004 ACM symposium on Applied computing. ACM Press, 439 - 446.
- [58] Sebyala, A. A., Olukemi, T., and Sacks, L. 2002. Active platform security through intrusion detection using naive Bayesian network for outlier detection. In Proceedings of the 2002 London Communications Symposium.
- [59] Valdes, A. and Skinner, K. 2000. Adaptive, model-based monitoring for cyber attack detection. In Proceedings of the 3rd International Workshop on Recent Advances in Intrusion Detection. Springer-Verlag, 80 - 92.
- [60] Bronstein, A., Das, J., Duro, M., Friedrich, R., Kleynner, G., Mueller, M., Singhal, S., and Cohen, I. 2001. Bayesian networks for detecting anomalies in internet based services. In International Symposium on Integrated Network Management
- [61] Barbara, D., Couto, J., Jajodia, S., and Wu, N. 2001a. Adam: a testbed for exploring the use of data mining in intrusion detection. *SIGMOD Rec.* 30, 4, 15 - 24.
- [62] Zhang, Z., Li, J., Manikopoulos, C., Jorgenson, J., and Ucles, J. 2001. HIDE: a hierarchical network intrusion detection system using statistical preprocessing and neural network classification. In Proceedings of IEEE Workshop on Information Assurance and Security. West Point, 85 - 90.
- [63] Labib, K. and Vemuri, R. 2002. NSOM: A real-time network-based intrusion detection using self-organizing maps. *Networks and Security*.
- [64] Smith, R., Bivens, A., Embrechts, M., Palagiri, C., and Szymanski, B. 2002. Clustering approaches for outlier based intrusion detection. In Proceedings of Intelligent Engineering Systems through Artificial Neural Networks. ASME Press, 579 - 584.
- [65] Williams, G., Baxter, R., He, H., Hawkins, S., and Gu, L. 2002. A comparative study of rnn for outlier detection in data mining. In Proceedings of the 2002 IEEE International Conference on Data Mining. IEEE Computer Society, Washington, DC, USA, 709.
- [66] Kruegel, C., Mutz, D., Robertson, W., and Valeur, F. 2003. Bayesian event classification for intrusion detection. In Proceedings of the 19th Annual Computer Security Applications Conference. IEEE Computer Society, 14.
- [67] Manikopoulos, C. and Papavassiliou, S. 2002. Network intrusion and fault detection: a statistical outlier approach. *IEEE Communication Magazine* 40.
- [68] Ramadas, M., Ostermann, S., and Tjaden, B. C. 2003. Detecting anomalous network traffic with self-organizing maps. In Proceedings of Recent Advances in Intrusion Detection. 36 - 54.
- [69] Ghosh, S. and Reilly, D. L. 1994. Credit card fraud detection with a neural-network. In Proceedings

- of the 27th Annual Hawaii International Conference on System Science. Vol. 3. Los Alamitos, CA.
- [70] Brause, R., Langsdorf, T., and Hepp, M. 1999. Neural data mining for credit card fraud detection. In Proceedings of IEEE International Conference on Tools with Artificial Intelligence. 103 - 106.
- [71] Dorransoro, J. R., Ginel, F., Sanchez, C., and Cruz, C. S. 1997. Neural fraud detection in credit card operations. IEEE Transactions On Neural Networks 8, 4 (July), 827 - 834.
- [72] Bolton, R. and Hand, D. 1999. Unsupervised profiling methods for fraud detection. In Credit Scoring and Credit Control VII.
- [73] Fawcett, T. and Provost, F. 1999. Activity monitoring: noticing interesting changes in behavior. In Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM Press, 53 - 62.
- [74] Cox, K. C., Eick, S. G., Wills, G. J., and Brachman, R. J. 1997. Visual data mining: Recognizing telephone calling fraud. Journal of Data Mining and Knowledge Discovery 1, 2, 225 - 231.
- [75] Aggarwal, C. 2005. On abnormality detection in spuriously populated data streams. In Proceedings of 5th SIAM Data Mining. 80 - 91.
- [76] Scott, S. L. 2001. Detecting network intrusion using a Markov modulated non homogeneous Poisson process. Submitted to the Journal of the American Statistical Association.
- [77] Barson, P., Davey, N., Field, S. D. H., Frank, R. J., and McAskie, G. 1996. The detection of fraud in mobile phone networks. Neural Network World 6, 4.
- [78] Taniguchi, M., Haft, M., Hollmn, J., and Tresp, V. 1998. Fraud detection in communications networks using neural and probabilistic methods. In Proceedings of IEEE International Conference in Acoustics, Speech and Signal Processing. Vol. 2. IEEE Computer Society, 1241 - 1244.
- [79] Phua, C., Alahakoon, D., and Lee, V. 2004. Minority report in fraud detection: classification of skewed data. SIGKDD Explorer Newsletter 6, 1, 50 - 59.
- [80] He, Z., Xu, X., and Deng, S. 2003. Discovering Cluster-based local outliers. Pattern Recognition Letters 24, 9-10, 1641 - 1650.
- [81] Brockett, P. L., Xia, X., and Derrig, R. A. 1998. Using Kohonen's self-organizing feature map to uncover automobile bodily injury claims fraud. Journal of Risk and Insurance 65, 2 (June), 245 - 274.
- [82] Donoho, S. 2004. Early detection of insider trading in option markets. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM Press, New York, NY, USA, 420 - 429.
- [83] Anscombe, F. J. and Guttman, I. 1960. Rejection of outliers. Technometrics 2, 2, 123 - 147. Arning, A., Agrawal, R., and Raghavan, P. 1996. A linear method for deviation detection in large databases. In Proceedings of 2nd International Conference of Knowledge Discovery and Data Mining. 164 - 169.
- [84] Horn, P. S., Feng, L., Li, Y., and Pesce, A. J. 2001. Effect of outliers and non healthy individuals on reference interval estimation. Clinical Chemistry 47, 12, 2137 - 2145.
- [85] Laurikkala, J., Juhola, M., and Kentala, E. 2000. Informal identification of outliers in medical data. In Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology. 20 - 24.
- [86] Solberg, H. E. and Lahti, A. 2005. Detection of outliers in reference distributions: Performance of Horn's algorithm. Clinical Chemistry 51, 12, 2326 - 2332.
- [87] Roberts, S. 1999. Novelty detection using extreme value statistics. In Proceedings of IEEE - Vision, Image and Signal processing. Vol. 146. 124 - 129.
- [88] Suzuki, E., Watanabe, T., Yokoi, H., and Takabayashi, K. 2003. Detecting interesting exceptions from medical test data with visual summarization. In Proceedings of the 3rd IEEE International Conference on Data Mining. 315 - 322.
- [89] Campbell, C. and Bennett, K. 2001. A linear programming approach to novelty detection. In Proceedings of Advances in Neural Information Processing. Vol. 14. Cambridge Press.
- [90] Wong, W.-K., Moore, A., Cooper, G., and Wagner, M. 2003. Bayesian network outlier pattern detection for disease outbreaks. In Proceedings of the 20th International Conference on Machine Learning. AAAI Press, Menlo Park, California, 808 - 815.
- [91] Lin, J., Keogh, E., Fu, A., and Herle, H. V. 2005. Approximations to magic: Finding unusual medical time series. In Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems. IEEE Computer Society, Washington, DC, USA, 329 - 334.
- [92] Guttormsson, S., II, R. M., and El Sharkawi, M. 1999. Elliptical novelty grouping for on-line short-turn detection of excited running rotors. IEEE Transactions on Energy Conversion 14, 1 (March).
- [93] Lin, J., Keogh, E., Fu, A., and Herle, H. V. 2005. Approximations to magic: Finding unusual medical time series. In Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems. IEEE Computer Society, Washington, DC, USA, 329 - 334.
- [94] Keogh, E., Lonardi, S., and Chi' Chiu, B. Y. 2002. Finding surprising patterns in a time series database in linear time and space. In Proceedings of the eighth ACM SIGKDD International conference on Knowledge discovery and data mining. ACM Press, New York, NY, USA, 550 - 556.
- [95] Keogh, E., Lin, J., Lee, S.-H., and Herle, H. V. 2006. Finding the most unusual time series subsequence: algorithms and applications. Knowledge and Information Systems 11, 1, 1 - 27.
- [96] Desforges, M., Jacob, P., and Cooper, J. 1998. Applications of probability density estimation to the detection of abnormal conditions in engineering. In Proceedings of Institute of Mechanical Engineers. Vol. 212. 687 - 703.
- [97] Bishop, C. 1994. Novelty detection and neural network validation. In Proceedings of IEEE Vision, Image and Signal Processing. Vol. 141. 217 - 222.
- [98] Diaz, I. and Hollmen, J. 2002. Residual generation and visualization for understanding novel process conditions.



- In Proceedings of IEEE International Joint Conference on Neural Networks. IEEE, Honolulu, HI, 2070 - 2075.
- [99] Harris, T. 1993. Neural network in machine health monitoring. Professional Engineering. Hartigan, J. A. and Wong, M. A. 1979. A k-means clustering algorithm. Applied Statistics 28, 100 - 108.
- [100] Jakubek, S. and Strasser, T. 2002. Fault-diagnosis using neural networks with ellipsoidal basis functions. In Proceedings of the American Control Conference. Vol. 5. 3846 - 3851.
- [101] King, S., King, D., P. Anuzis, K. A., Tarassenko, L., Hayton, P., and Utete, S. 2002. The use of novelty detection techniques for monitoring high-integrity plant. In Proceedings of the 2002 International Conference on Control Applications. Vol. 1. Cancun, Mexico, 221 - 226.
- [102] Li, Y., Pont, M. J., and Jones, N. B. 2002. Improving the performance of radial basis function classifiers in condition monitoring and fault diagnosis applications where unknown faults may occur. Pattern Recognition Letters 23, 5, 569 - 577.
- [103] Petsche, T., Marcantonio, A., Darken, C., Hanson, S., Kuhn, G., and Santoso, I. 1996. A neural network auto associator for induction motor failure prediction. In Proceedings of Advances in Neural Information Processing. Vol. 8. 924 - 930.
- [104] Streifel, R., Maks, R., and El-Sharkawi, M. 1996. Detection of shorted-turns in the field of turbine-generator rotors using novelty detectors - development and field tests. IEEE Transactions on Energy Conversations 11, 2, 312 - 317.
- [105] Whitehead, B. and Hoyt, W. 1993. A function approximation approach to outlier detection in propulsion system test data. In Proceedings of 29th AIAA/SAE/ASME/ASEE Joint Propulsion Conference. IEEE Computer Society, Monterey, CA, USA.
- [106] Parra, L., Deco, G., and Miesbach, S. 1996. Statistical independence and novelty detection with information preserving nonlinear maps. Neural Computing 8, 2, 260 - 269.
- [107] Yairi, T., Kato, Y., and Hori, K. 2001. Fault detection by mining association rules from house-keeping data. In Proceedings of International Symposium on Artificial Intelligence, Robotics and Automation in Space.
- [108] Manson, G. 2002. Identifying damage sensitive, environment insensitive features for damage detection. In Proceedings of the IES Conference. Swansea, UK.
- [109] Manson, G., Pierce, G., and Worden, K. 2001. On the long-term stability of normal condition for damage detection in a composite panel. In Proceedings of the 4th International Conference on Damage Assessment of Structures. Cardiff, UK.
- [110] Manson, G., Pierce, S. G., Worden, K., Monnier, T., Guy, P., and Atherton, K. 2000. Long-term stability of normal condition data for novelty detection. In Proceedings of Smart Structures and Integrated Systems. 323 - 334.
- [111] Ruotolo, R. and Surace, C. 1997. A statistical approach to damage detection through vibration monitoring. In Proceedings of the 5th Pan American Congress of Applied Mechanics. Puerto Rico.
- [112] Hickinbotham, S. J. and Austin, J. 2000a. Novelty detection in airframe strain data. In Proceedings of 15th International Conference on Pattern Recognition. Vol. 2. 536 - 539.
- [113] Hickinbotham, S. J. and Austin, J. 2000b. Novelty detection in airframe strain data. In Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. Vol. 6. 24 - 27.
- [114] Hollier, G. and Austin, J. 2002. Novelty detection for strain-gauge degradation using maximally correlated components. In Proceedings of the European Symposium on Artificial Neural Networks. 257 - 262 - 539.
- [115] Brotherton, T. and Johnson, T. 2001. Outlier detection for advance military aircraft using neural networks. In Proceedings of 2001 IEEE Aerospace Conference.
- [116] Brotherton, T., Johnson, T., and Chadderdon, G. 1998. Classification and novelty detection using linear models and a class dependent - elliptical basis function neural network. In Proceedings of the IJCNN Conference. Anchorage AL.
- [117] Nairac, A., Corbett-Clark, T., Ripley, R., Townsend, N., and Tarassenko, L. 1997. Choosing an appropriate model for novelty detection. In Proceedings of the 5th IEEE International Conference on Artificial Neural Networks. 227 - 232.
- [118] Nairac, A., Townsend, N., Carr, R., King, S., Cowley, P., and Tarassenko, L. 1999. A system for the analysis of jet engine vibration data. Integrated Computer-Aided Engineering 6, 1, 53 - 56.
- [119] Surace, C. and Worden, K. 1998. A novelty detection method to diagnose damage in structures: an application to an offshore platform. In Proceedings of Eighth International Conference of Off-shore and Polar Engineering. Vol. 4. Colorado, USA, 64 - 70.
- [120] Surace, C., Worden, K., and Tomlinson, G. 1997. A novelty detection approach to diagnose damage in a cracked beam. In Proceedings of SPIE. Vol. 3089. 947 - 953.
- [121] Sohn, H., Worden, K., and Farrar, C. 2001. Novelty detection under changing environmental conditions. In Proceedings of Eighth Annual SPIE International Symposium on Smart Structures and Materials. Newport Beach, CA.
- [122] Worden, K. 1997. Structural fault detection using a novelty measure. Journal of Sound Vibration 201, 1, 85 - 101.
- [123] Augustejn, M. and Folkert, B. 2002. Neural network classification and novelty detection. International Journal on Remote Sensing 23, 14, 2891 - 2902.
- [124] Byers, S. D. and Raftery, A. E. 1998. Nearest neighbor clutter removal for estimating features in spatial point processes. Journal of the American Statistical Association 93, 577 - 584.
- [125] Moya, M., Koch, M., and Hostetler, L. 1993. One-class classifier networks for target recognition applications. In Proceedings on World Congress on Neural Networks, International Neural Network Society. Portland, OR, 797 - 801.

- [126] Torr, P. and Murray, D. 1993. Outlier detection and motion segmentation. In Proceedings of SPIE, Sensor Fusion VI, Paul S. Schenker; Ed. Vol. 2059. 432 - 443.
- [127] Theiler, J. and Cai, D. M. 2003. Re-sampling approach for outlier detection in multispectral images. In Proceedings of SPIE 5093, 230-240, Ed.
- [128] Cun, Y. L., Boser, B., Denker, J. S., Howard, R. E., Hubbard, W., Jackel, L. D., and Henderson, D. 1990. Handwritten digit recognition with a back-propagation network. Advances in neural information processing systems, 396 - 404.
- [129] Spence, C., Parra, L., and Sajda, P. 2001. Detection, synthesis and compression in mammographic image analysis with a hierarchical image probability model. In Proceedings of the IEEE Workshop on Mathematical Methods in Biomedical Image Analysis. IEEE Computer Society, Washington, DC, USA, 3.
- [130] Tarassenko, L. 1995. Novelty detection for the identification of masses in mammograms. In Proceedings of the 4th IEEE International Conference on Artificial Neural Networks. Vol. 4. Cambridge, UK, 442 - 447.
- [131] Chen, D., Shao, X., Hu, B., and Su, Q. 2005. Simultaneous wavelength selection and outlier detection in multivariate regression of near-infrared spectra. Analytical Sciences 21, 2, 161 - 167.
- [132] Davy, M. and Godsill, S. 2002. Detection of abrupt spectral changes using support vector machines. an application to audio signal segmentation. In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Orlando, USA.
- [133] Hazel, G. G. 2000. Multivariate Gaussian MRF for multispectral scene segmentation and outlier detection. GeRS 38, 3 (May), 1199 - 1211.
- [134] Scarth, G., McIntyre, M., Wowk, B., and Somorjai, R. 1995. Detection of novelty in functional images using fuzzy clustering. In Proceedings of the 3rd Meeting of International Society for Magnetic Resonance in Medicine. Nice, France, 238.
- [135] Diehl, C. and Hampshire, J. 2002. Real-time object classification and novelty detection for collaborative video surveillance. In Proceedings of IEEE International Joint Conference on Neural Networks. IEEE, Honolulu, HI.
- [136] Singh, S. and Markou, M. 2004. An approach to novelty detection applied to the classification of image regions. IEEE Transactions on Knowledge and Data Engineering 16, 4, 396 - 407. To Appear in ACM Computing Surveys, 09 2009.
- [137] Pokrajac, D., Lazarevic, A., and Latecki, L. J. 2007. Incremental local outlier detection for data streams. In Proceedings of IEEE Symposium on Computational Intelligence and Data Mining.
- [138] Song, Q., Hu, W., and Xie, W. 2002. Robust support vector machine with bullet hole image classification. IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews 32, 4.
- [139] Baker, D., Hofmann, T., McCallum, A., and Yang, Y. 1999. A hierarchical probabilistic model for novelty detection in text. In Proceedings of International Conference on Machine Learning.
- [140] Manevitz, L. M. and Yousef, M. 2000. Learning from positive data for document classification using neural networks. In Proceedings of Second Bar-Ilan Workshop on Knowledge Discovery and Learning. Jerusalem.
- [141] Manevitz, L. M. and Yousef, M. 2002. One-class SVMs for document classification. Journal of Machine Learning Research 2, 139 - 154.
- [142] Allan, J., Carbonell, J., Doddington, G., Yamron, J., and Yang, Y. 1998. Topic detection and tracking pilot study. In Proceedings of DARPA Broadcast News Transcription and Understanding Workshop. 194 - 218.
- [143] Srivastava, A. and Zane-Ulman, B. 2005. Discovering recurring outliers in text reports regarding complex space systems. Aerospace Conference, 2005 IEEE, 3853 - 3862.
- [144] Srivastava, A. 2006. Enabling the discovery of recurring outliers in aerospace problem reports using high-dimensional clustering techniques. Aerospace Conference, 2006 IEEE, 17 - 34.
- [145] Chatzigiannakis, V., Papavassiliou, S., Grammatikou, M., and Maglaris, B. 2006. Hierarchical outlier detection in distributed large-scale sensor networks. In ISCC '06: Proceedings of the 11th IEEE Symposium on Computers and Communications. IEEE Computer Society, Washington, DC, USA, 761 - 767.
- [146] Janakiram, D., Reddy, V., and Kumar, A. 2006. Outlier detection in wireless sensor networks using Bayesian belief networks. In First International Conference on Communication System Software and Middleware. 1 - 6.
- [147] Branch, J., Szymanski, B., Giannella, C., Wolff, R., and Kargupta, H. 2006. In-network outlier detection in wireless sensor networks. In 26th IEEE International Conference on Distributed Computing Systems.
- [148] Phuong, T. V., Hung, L. X., Cho, S. J., Lee, Y., and Lee, S. 2006. An outlier detection algorithm for detecting attacks in wireless sensor networks. Intelligence and Security Informatics 3975, 735 - 736.
- [149] Du, W., Fang, L., and Peng, N. 2006. Lad: localization outlier detection for wireless sensor networks. J. Parallel Distrib. Comput. 66, 7, 874 - 886.
- [150] Subramaniam, S., Palpanas, T., Papadopoulos, D., Kalogeraki, V., and Gunopulos, D. 2006. Online outlier detection in sensor data using non-parametric models. In VLDB '06: Proceedings of the 32nd international conference on Very large data bases. VLDB Endowment, 187 - 198.
- [151] Kejia Zhang, Shengfei Shi, H. G. and Li, J. 2007. Unsupervised outlier detection in sensor networks using aggregation tree. Advanced Data Mining and Applications 4632, 158 - 169.
- [152] Ide, T., Papadimitriou, S., and Vlachos, M. 2007. Computing correlation outlier scores using stochastic nearest neighbors. In Proceedings of International Conference Data Mining. 523 - 528.
- [153] Albrecht, S., Busch, J., Kloppenburg, M., Metze, F., and Tavan, P. 2000. Generalized radial basis function networks for classification and novelty detection: self-organization of optional Bayesian decision. Neural Networks 13, 10, 1075 - 1093.
- [154] Emamian, V., Kaveh, M., and Tewfik, A. 2000. Robust clustering of acoustic emission signals using the Kohonen

- network. In Proceedings of the IEEE International Conference of Acoustics, Speech and Signal Processing. IEEE Computer Society.
- [155] Crook, P. and Hayes, G. 2001. A robot implementation of a biologically inspired method for novelty detection. In Proceedings of Towards Intelligent Mobile Robots Conference. Manchester, UK.
- [156] Crook, P. A., Marsland, S., Hayes, G., and Nehmzow, U. 2002. A tale of two filters - on-line novelty detection. In Proceedings of International Conference on Robotics and Automation. 3894 - 3899.
- [157] Marsland, S., Nehmzow, U., and Shapiro, J. 1999. A model of habituation applied to mobile robots. In Proceedings of Towards Intelligent Mobile Robots. Department of Computer Science, Manchester University, Technical Report Series, ISSN 1361-6161, Report UMCS-99-3-1.
- [158] Marsland, S., Nehmzow, U., and Shapiro, J. 2000b. A real-time novelty detector for a mobile robot. In Proceedings of the EUREL Conference on Advanced Robotics Systems.
- [159] Marsland, S., Nehmzow, U., and Shapiro, J. 2000a. Novelty detection for robot neotaxis. In Proceedings of the 2nd International Symposium on Neural Computation. 554 - 559.
- [160] Shekhar, S., Lu, C.-T., and Zhang, P. 2001. Detecting graph-based spatial outliers: algorithms and applications (a summary of results). In Proceedings of the 7th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM Press, New York, NY, USA, 371 - 376.
- [161] Ihler, A., Hutchins, J., and Smyth, P. 2006. Adaptive event detection with time-varying poisson processes. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM Press, New York, NY, USA, 207 - 216.
- [162] Ide, T. and Kashima, H. 2004. Eigenspace-based outlier detection in computer systems. In Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM Press, New York, NY, USA, 440 - 449.
- [163] Sun, J., Qu, H., Chakrabarti, D., and Faloutsos, C. 2005. Neighborhood formation and outlier detection in bipartite graphs. In Proceedings of the 5th IEEE International Conference on Data Mining. IEEE Computer Society, Washington, DC, USA, 418 - 425. To Appear in ACM Computing Surveys, 09 2009.
- [164] MacDonald, J. W. and Ghosh, D. 2007. Copa - cancer outlier profile analysis. *Bioinformatics* 22, 23, 2950 - 2951.
- [165] Kadota, K., Tominaga, D., Akiyama, Y., and Takahashi, K. 2003. Detecting outlying samples in microarray data: A critical assessment of the effect of outliers on sample classification. *Chem-Bio Informatics* 3, 1, 30 - 45.
- [166] Tomlins, S. A., Rhodes, D. R., Perner, S., Dhanasekaran, S. M., Mehra, R., Sun, X. W., Varambally, S., Cao, X., Tchinda, J., Kuefer, R., Lee, C., Montie, J. E., Shah, R., Pienta, K. J., Rubin, M., and Chinnaiyan, A. M. 2005. Recurrent fusion of *tmprss2* and *ETS* transcription factor genes in prostate cancer. *Science* 310, 5748, 603 - 611.
- [167] Tibshirani, R. and Hastie, T. 2007. Outlier sums for differential gene expression analysis. *Biostatistics* 8, 1, 2 - 8.
- [168] Lu, C.-T., Chen, D., and Kou, Y. 2003. Algorithms for spatial outlier detection. In Proceedings of 3rd International Conference on Data Mining. 597 - 600.
- [169] Lin, S. and Brown, D. E. 2003. An outlier-based data association method for linking criminal incidents. In Proceedings of 3rd SIAM Data Mining Conference.
- [170] He, Z., Xu, X., Huang, J. Z., and Deng, S. 2004b. Mining class outliers: Concepts, algorithms and applications. 588 - 589.
- [171] Dutta, H., Giannella, C., Borne, K., and Kargupta, H. 2007. Distributed top-k outlier detection in astronomy catalogs using the DEMAC system. In Proceedings of 7th SIAM International Conference on Data Mining.
- [172] Escalante, H. J. 2005. A comparison of outlier detection algorithms for machine learning. In Proceedings of the International Conference on Communications in Computing.
- [173] Protopoulos, P., Giammarco, J. M., Faccioli, L., Struble, M. F., Dave, R., and Alcock, C. 2006. Finding outlier light curves in catalogues of periodic variable stars. *Monthly Notices of the Royal Astronomical Society* 369, 2, 677 - 696.
- [174] Blender, R., Fraedrich, K., and Lunkeit, F. 1997. Identification of cyclone-track regimes in the North Atlantic. *Quarterly Journal of the Royal Meteorological Society* 123, 539, 727 - 741.
- [175] Sun, P. and Chawla, S. 2004. On local spatial outliers. In Proceedings of 4th IEEE International Conference on Data Mining. 209 - 216.
- [176] Sun, P., Chawla, S., and Arunasalam, B. 2006. Mining for outliers in sequential databases. In SIAM International Conference on Data Mining.

# Pair of Iris Recognition for Personal Identification Using Artificial Neural Networks

K.Saminathan<sup>1</sup> M.Chithra Devi<sup>2</sup> T.Chakravarthy<sup>3</sup>

<sup>1</sup> Assistant Professor, Prist University, Kumbakonam, Tamil Nadu, India.

<sup>2</sup> Assistant Professor, Periyar Maniammai University, Thanjavur, Tamil Nadu, India.

<sup>3</sup> Associate Professor, A.V.V.M. Sri Pushpam College, Poondi, Tamil Nadu, India.

## Abstract

Pair of iris recognition is very effective for person identification due to the iris unique features and the protection of the iris from the environment and aging. In addition it is well suitable to embark upon accidental or ophthalmological disease issue. This paper presents a simple methodology for pre-processing pair of iris images which means both left and right eye of human (instead of either right or left eye) and the design and training of feedforward artificial neural network for iris recognition system. Three different iris image data partitioning techniques and two data coding are proposed and explored. We also experiment with various number of hidden layers, number of neurons in each hidden layer, input format (binary vs. analog) percent of data used for training vs testing, and with the addition of noise. Our recognition system achieves high accuracy despite using simple data preprocessing and a simple neural network.

## Keywords:

*Pair of iris recognition, feedforward neural networks, Backpropagation training algorithm, Pre-processing, data partitioning.*

## 1. Introduction

Biometric measures [1] such as recognizing one's fingerprints, face, iris and voice greatly help in person identification authentication, and authorization. Pair of iris recognition has the high potential and noninvasive personal verification. This is because each person's iris is unique even for twins and hardly changes while other biometric measures are quite intrusive to the operator and offer little possibility of covert evaluation. In fact, the eyelid, cornea and aqueous humor protect the iris. Furthermore, the iris is relatively immune to aging, and the wearing of contact lenses or glasses. In addition to that pair of iris recognition system well suitable to embark upon accidental and ophthalmological issue. Therefore, pair of iris recognition is a biometric technique which can be trusted for producing accurate and correct results. As irises differ in size, shape, color and patterns, they offer high confidence for recognizing a person's identity by mathematical analysis. Typically, iris recognition software based on advanced mathematical techniques such as wavelets and real-time hardware with high resolution camera(s) is employed. In this work we take a different approach aiming at

simplifying the iris recognition system with high accuracy.

Our approach is characterized by (i) the use of artificial neural networks and (ii) simple mathematical analysis and pre-processing the pair of iris. The paper presents a methodology for pre-processing iris images and getting them ready for artificial neural network processing. The paper also presents the design and training of an artificial neural network for recognizing several pair of iris.

## 2. Related and Background Work

### 2.1 Iris Recognition

The iris patterns have a wonderful and rich structure and are full of complex textures in contrast to other physiological characteristics. The iris is the colored circular part of the eye between the pupil and sclera. When security is highly desired, iris recognition is a preferred method of identification because the iris patterns are hardly affected by the environment and can hardly be lost. The iris texture is unique from one person to another. Some of the iris features include furrows, ridges, arching ligaments, zigzag collarets, and crypts.

We next review relevant background work, Daugman [2] used 2D Gabor filters and phase coding to generate a 2048 binary feature code for the iris. Wildes [3] used the Hough transform to locate the iris and a Laplacian pyramid with four resolution levels to generate the iris code. Boles and Boashash [4] built a 1D representation of the grey level signature of the iris and applied to it zero-crossing of the dyadic wavelet transform to generate the iris representation. Wang and Tan [5] used a bank of Gabor filters to capture the iris profile. Based on a 2D Haar wavelet, extracted high frequency information to generate an 87 binary code and employed an LVQ neural network for classification. Zhang and ma [6] employed an intersecting cortical model (ICM) neural network to generate the iris codes and the Hamming distance between the compared iris codes. This ICM neural network is a simplified model of pulse – coupled neural network (PCNN), which has excellent performance for image segmentation, so the coding process is fast enough.

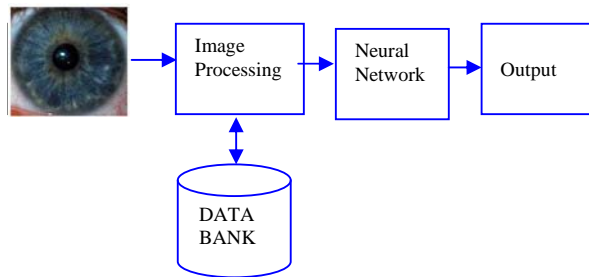


Fig. 1. Block Diagram of iris recognition.

## 2.2 Other Biometric Techniques

A biometric system is essentially a pattern recognition system that operates by acquiring physiological and / or behavioral characteristic data from a person, extracting some features from the acquired data, and comparing these features against a recorded feature set in the database [8] for the purpose of determining or confirming the person's identity. Biometric applications include computer systems security, secure electronic banking, mobile phones, credit cards, secure access to buildings, health and social services. By using biometrics a person could be identified based on her / his Physiological and / or identity rather than her / his possession (card, token, key) or her / his knowledge (e.g. password, PIN).

Desirable characteristics of a biometric recognition system include (i) universality: The feature should apply to every person or special alternative tests should be administered to those who do not apply, e.g. blind or person without fingerprints (ii) uniqueness: the system should extract and compare a feature unique to each person (iii) longevity: the feature should not vary with time, (iv) collectability: the feature must be easily collectible (v) accuracy: the system should deliver accurate recognition, and (vi) tampering: the technique should be hard to tamper.

## 2.3 Artificial Neural Networks

Artificial neural networks model biological neural networks in the brain and have proven their effectiveness in a number of applications such as classification and categorization, prediction, pattern recognition and control. An artificial neural network consists of interconnected groups of artificial neurons. Such a network performs computation and manipulates information based on the connectionist approach in a similar but simpler fashion than the brain would perform. Many types of artificial neural networks [7] exist including feed forward neural networks, radial basis function (RBF) networks, Kohonen self-organizing networks, recurrent networks, stochastic neural networks, modular neural networks, dynamic neural networks, cascading neural networks, and fuzzy neuro networks. Multi-layer perception [8] (MLP) is perhaps the most popular, where neurons in a feedforward type network perform a biased weighted

averaging of their inputs and this sum is then subjected to a transfer function, in order to limit the output value. The MLP is an example of feedforward artificial neural network with multiple layers and where each neuron output in one layer feeds as input to the neurons in the next layer as shown in fig. (2).

We chose our artificial neural network for iris recognition of the feedforward type due to its simplicity and its suitability for this application.

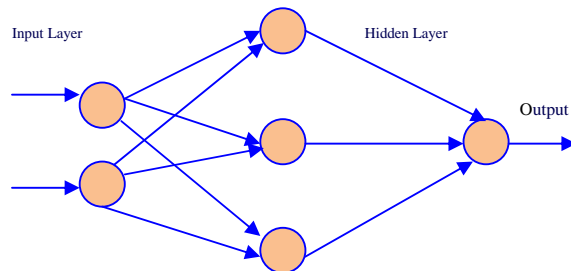


Fig. 2. MLP neural network

We also employ the back propagation algorithm for supervised training of our network, a well known and widely used algorithm. The training algorithm minimizes the error between the obtained output and the required target output by finding the lowest point or minimum in the Error surface. Starting with initial weights and thresholds, the training algorithms look for the global minimum of the error surface. Usually, the slope of the error surface, at the present location and guides the next move down.

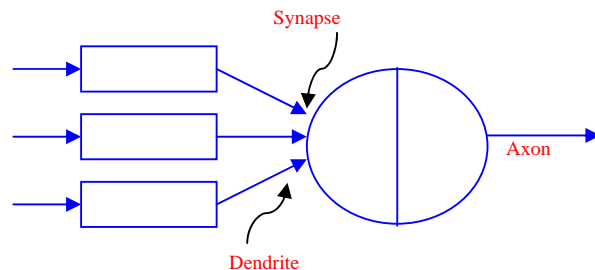


Fig. 3. Artificial neuron model

An artificial neuron models a real neuron as depicted in fig.(3). First, electric signals from other neurons are multiplied by weights (represented by the rectangles in fig.(2) and then are input into the artificial neuron. The weighted signal values are then summed by an adder ("Σ" in fig.2) and the sum is subjected to a transfer function ("T" in fig.2) which is one of : (i) linear, where the output is proportional to the weighted sum of inputs; (ii) threshold, where the output is one of two values based on whether the weighted sum is greater or smaller than the threshold value; and (iii) sigmoid, a non – linear function which most closely mimics real neurons. Artificial neural networks are composed of several artificial neurons as a real neuron network is composed of many real neurons. Artificial neural networks come in different forms and shapes.

### 3. Proposed Work

#### 3.1 Preprocessing

To implement the pair of iris recognition system we gathered pair of iris images for training and testing the neural network, we decided to select iris images of the same color (brown) in order to create more difficult situations for our recognition system to detect and achieve higher recognition accuracy. Up close, the irises that have been collected are different in their patterns and shapes although from a further distance, the irises images look similar to each other. We collected and pre-processed 20 brown colored pair of iris images of different persons(both left and right) from the iris database (Chek image database). The iris images were between 400 KB and 500 KB and were not ready for processing but had to be pre-processed. For instance the white Sclera and black pupil are visible in all the images. Additionally, the relevant content of the binary iris image was not ready to be fed to an artificial neural network for processing.

In our work we manually manipulated the images as our focus was on the design of artificial neural network by using Adobe Photoshop, Java program(Lin,xxx) and Excel spread sheet to retrieve the required iris images with the exclude of sclera and pupil. The DAT file as input to Brain Maker simulator obtained from Netmaker application.

#### 3.2. Iris Image Data Partitioning

As it is desired to reduce the cost of the artificial neural network, and as Brain maker limits the number of neurons per layer, the iris image's RGB matrix had to be partitioned to reduce as much possible the number of values fed as input to the neural network. For that purpose, we considered three simple data partitioning techniques pictured in fig.4.

1. Horizontal strip partitioning (rows)
2. Vertical strip partitioning (columns)
3. Block partitioning.

In horizontal strip (row) partitioning we divided the RGB matrix into  $r=(10, 25)$  horizontal strips and summed all the RGB values of all pixels falling in one horizontal strip, when  $r$  was set to 10 or 25, and as the image contains 100 x 100 pixels, each horizontal strip, contained the RGB values of  $100/10(25)= 10$  (4) rows of 100 pixels each, and a 1000 (400) RGB values were summed into one number representing that horizontal strip.



Fig. 4. Data partitioning techniques: horizontal (left), vertical (middle), block (right).

Similarly, in vertical strip (column) partitioning, we divided the RGB matrix into  $c = (10, 25)$  vertical strips and summed all the RGB values of all pixels falling in one vertical strip. When  $C$  was set to 10 (25), and as the image contains 100 x 100 pixels, each vertical strip contained the RGB values of  $100 / 10 (25) = 10$  (4) columns of 100 pixels each, and 1000 (400) RGB values were summed into one number representing that vertical strip. In block partitioning, the image was divided into  $b = (16, 25)$  square blocks. When  $b$  was set to 16 (25), and as the image contains 100 x 100 pixels, each block consists of 25 x 25 (20 x 20) pixels whose RGB values were summed into one number per block.

Each horizontal / vertical strip or block was thus represented by one number which was fed as one input into the artificial neural network for identifying a match / no match of the presented iris image. Our data partitioning techniques are characterized by simplicity and fast processing, compared to more complicated techniques based on wavelets, and are key to reduce the system cost.

### 4. Performance Issues

As the NetMaker accepts DAT files as inputs, these DAT, files containing the sums of RGB values in the various strips and blocks previously discussed were prepared for that purpose. Neural network training and testing experiments were conducted for two different data encodings: binary and analog. In binary coding, each sum is converted into 6 bits for both horizontal and vertical strip partitioning and 4 bits for block partitioning. These numbers are governed by the maximum number of neurons per layer (=64) acceptable by the BrainMaker Simulator.

Table 1: Experiments results with 10 horizontal or vertical strips, or 16 blocks.

Experiment description	Incorrect detection	Accuracy (%)
(Rows) Analog input: with one hidden layer (10 neurons)	5/15	66.67
(Columns) Analog input: with one hidden layer (10 neurons)	5/15	66.67
(Blocks) Analog input: with one hidden layer (10 neurons)	2/15	86.67
(Blocks) Analog input: with one hidden layer (50 neurons)	1/15	93.34
(Blocks) Analog input: with one hidden layer (10 neurons each)	2/15	86.67
(Blocks) Analog input: with one hidden layer (5 and 10 neurons)	2/15	86.67
(Rows) Binary input: with one hidden layer (50 neurons) and without noise	10/15	33.33
(Columns) Binary input: with one hidden layer (50 neurons)	10/15	33.33
(Blocks) Analog input: with one hidden layer (64 neurons)	10/15	33.33

Table 2: Experiments results with 10 horizontal or vertical strips, or 16 blocks.

Experiment description	Incorrect detection	Accuracy (%)
(Rows) Analog input: with one hidden layer (25 neurons)	6/15	60
(Columns) Analog input: with one hidden layer (25 neurons)	6/15	60
(Blocks) Analog input: with one hidden layer (25 neurons)	4/15	73.33
(Blocks) Analog input: with one hidden layer (25 neurons)	6/15	60
(Blocks) Analog input: with one hidden layer (25 neurons each)	7/15	53.33
(Blocks) Analog input: with one hidden layer (25 and 50 neurons)	7/15	53.33

The best accuracy (93.34%) was obtained with 10 block partitioning with 10 neurons in the input layer and 50 neurons in the hidden layer and 1 hidden layer only. When the number of neurons in the hidden layer was reduced to 10 neurons to match the number of neurons in the input layer, the accuracy dropped to 86.66%. This accuracy result was obtained with 1 or 2 hidden layer of 10 neurons each, or with the first hidden layer containing 5 neurons and the second hidden layer containing 10 neurons. Also increasing the number of hidden layers from 1 to 2 reduced the accuracy. Another advantage of our approach is that our neural network directly issues a match or no match output while in other's work, a neural network computer an iris code which must later be subjected to a Hamming distance computation to indicate a match.

## 5. Conclusion

Pair of iris recognition is an efficient biometric method for personal identification and verification, which provides more security due to iris unique and plentiful complex structure by its nature. In this paper we addressed the problem pair of iris recognition using a simple feedforward artificial neural network trained with the backpropagation algorithm. We described a pre-processing method to prepare the neural network inputs from the pair of iris images. Our approach uses simple RGB value summing in each partition and a simple MLP feed-forward neural network and issues a match/no match result without having to subject iris codes to Hamming distances. In this paper block partitioning results high performance and accuracy of 93.34% was obtained than the other two partitioning methods. All these features lead to high performance of the biometric system to provide high security. In future, the implementation system cost can be reduced by using single neural network and to support various colors of iris patterns.

## References

[1] Delac, K., & Grgic.M. A survey of biometric recognition methods. In 46<sup>th</sup> international

symposium electronics in marine, ELMAR – 2004, zadal croatia.

[2] Daugman .J, How iris recognition works. In proceedings of international conference on image processing (vol.1), 2002.

[3] Wildes, R, A system for automated iris recognition. In proceedings of 2<sup>nd</sup> IEEE work shop on applications of computer vision, 1994.

[4] Boles. W., & Boashash.B, A human identification technique using images of the iris and wave let transform. IEEE transactions on signal processing, 46 (4) 2008.

[5] Ma.L., wang. Y., & Tan. T, Iris recognition using circular symmetric filters. In proceedings of 16<sup>th</sup> international conference pattern recognition, 2009.

[6] Xu. G., zhang. Z., & Ma. Y, An efficient iris recognition system based on intersecting cortical model neural network. International Journal of cognitive informatics and natural Intelligence, 2008.

[7] Jain.A., Ross.A., & Prabhakar.S, An introduction to biometric recognition. IEEE transactions on circuits and systems for video Technology. Special Issue on Image and video – Based Biometrics, 2007.

[8] Haykin.S, Neural networks: A comprehensive foundation (2<sup>nd</sup> ed). NY: Prentice Hall, 2008.



Mr. K.Saminathan received M.Sc., M.Phil., M.Tech., degree in Computer Science and Engineering. He is currently doing Ph.D degree in Computer science in Bharathidasan University, Tiruchirappalli. He is working as Assistant Professor in Department of Computer science and Engineering at PRIST University, Kumbakonam(Dt), Tamil Nadu, India. He

has presented papers in international conferences and published papers in international journals.



Ms. M.Chithra Devi received M.Sc., M.Phil., degree in Computer Science. She is currently doing M.E degree in Computer Science and Engineering. She is working as Assistant Professor in Department of Software Engineering at Periyar Maniammai University, Thanjavur(Dt), Tamil Nadu, India. She has presented papers in international and national conferences and her research

area is image processing.



Dr. T.Chakravarthy awarded Ph.D., degree in Computer Science at Bharathidasan University, Tiruchirappalli. He is currently working as Associate Professor, Department of Computer Science A.V.V.M Sri Pushpam College (Autonomous), Poondi, Thanjavur(Dt), Tamil Nadu, India. He has finished minor research project at UGC in India. He has presented several papers in international conferences and

international journals.

# A New Approach to the Data Aggregation in Wireless Sensor Networks

Chamran Asgari

Department of Computer Engineering, Islamic Azad University, Arak Branch, Arak, Iran

Javad Akbari Torkestani

Young Researchers Club, Arak Branch, Islamic Azad University, Arak, Iran

## Abstract

Several algorithms have been developed for problems of data aggregation in wireless sensor networks, all of which tried to increase networks lifetime. In this paper, we deal with this problem using a more efficient method, and offer a heuristic algorithm based on distributed learning automata to solve data aggregation problems within stochastic graphs. Given that data aggregating through creating backbones and making connected dominating sets (CDS) in networks lowers the ratio of responding hosts to the hosts existing in virtual backbones, we employed this idea to our algorithm, trying to increase networks lifetime considering such parameters as sensors lifetime, remaining and consumption energies in order to have an almost optimal data aggregation within networks. Finally, we assess our algorithm for make CDS lifetime given increased transmission range and increased sensors number.

**Keywords:** *Wireless sensor network, Data aggregation, Connected Dominating Set, Backbone formation, distributed learning automata.*

## 1. Introduction

Wireless sensor networks consist of a large number of inexpensive sensor nodes distributed in environment uniformly, having limited energy, therefore, in the most cases, nodes communicate with central node via their neighbors [1]. On the other hand, an optimal route must be selected because there are different routes to central node from any other nodes. On the other hand, frequent use of one route results in reduction of energy of sensors located on that route and, ultimately, in sensors destruction. For solving this problem, we can consider a wireless sensor network as a graph in the nodes (hosts) which are the sensors and edges show the links between sensors. If a backbone can be created in this graph the constituent nodes that are able to communicate with all graph nodes or, in other words, to cover them, it is not necessary to use all graph nodes to aggregate the data and it suffices only to carry out data aggregation on backbone nodes, then, to send the result in the form of a single packet to central node. The set of nodes

constituting backbone are referred to Connected Dominating Set (CDS) and each node of this set is called dominator. Creating CDS to aggregate data is a promising approach for reducing routing overhead since messages are transmitted only within virtual backbone by means of CDS and, also, data aggregating through lowers the ratio of responding hosts to the hosts existing in virtual backbones [2-5]. By offering an intellectual algorithm, we tried to increase networks lifetime considering such parameters as sensors lifetime, remaining and consumed energies of sensors, in order to have an almost optimal data aggregation within networks. Our algorithm operates as follows: initially, wireless sensor network is modeled as a unit disk graph  $G=(V, E)$  in the nodes that represent hosts and edges show the links among hosts [6, 7]; then, an intellectual algorithm based on distributed learning automata is implemented on the model to aggregate data. For this algorithm, each host is equipped with a learning automaton. Sink node is considered as the first dominator here. Next, learning automata selects next action randomly from its variable action set with respect to action probability vector and this process continues until finally entire network is covered, with set of selected dominators constituting the backbone. After that, message "data aggregation" is sent to dominators from sink node inside backbone. Dominators will send the message to their parents immediately after receiving it. Each parent must wait until it receives data from all its children, then, aggregates all data received from its children and sends it to its own parent until aggregates data is sent to sink node in the form of a single packet.

Once every iteration of the process has finished, action probability vectors are updated for any learning automata. Eventually, with iteration of process, learning automata converges to public policy of optimal data aggregation for network graph. Given that lifetime of created CDS is of special importance, the algorithm will pursue the aim of choosing a CDS with the longest lifetime from made CDSs.

The rest of this paper consists of five sections. The related work is reviewed in the section 2. In the section 3, some definitions and primary concepts will



be presented. In the Section 4, the proposed DLA-based backbone formation algorithm for finding a CDS with longest lifetime is presented. The experiment results are demonstrated in the section 5 and finally are in the section 6, the conclusion and future work is highlighted.

## 2. Related work

Many routing algorithms have been provided for the sensor networks. For some of these algorithms, each node may have more than one route to sink node that one of them is selected on the basis of a series of criteria, among the level of energy consumption along the route can be a proper criterion. Energy saving can be taken into account in two ways: (1) energy consumption is calculated for any separate routes, then, the route with minimal energy consumption is chosen [8]; and (2) data aggregation is based on provided learning automata, which prevents extra packets from being sent in networks by identifying sensors generating identical data and by activating sensor nodes periodically, and saves a large amount of energy while increasing network lifetime [9]. A solution has been provided in [10] for data aggregating and routing with intra network aggregations in wireless sensor networks in order to maximize network lifetime by using intra network processing techniques and data aggregation. The relationship between the security and data aggregation process within wireless sensor networks has been investigated in [11].

In [12], network is first clustered in order to aggregate data, then, head-clusters aggregate data from each cluster separately. A network organized into clusters with the same sizes results in unequal load distribution among head-cluster nodes. But [12] provides a model in which clusters are of different sizes, resulting in more uniform energy distribution among head-cluster nodes and with increasing in network lifetime. In [13] has offered data aggregation in wireless sensor networks by using ant colony algorithm that states the problem of creating data aggregation tree in wireless sensor networks for a group of source nodes to send sensed data to the single sink node. Ant colony system represents a natural method of heuristic search to determining data aggregation. Each ant discovers all possible routes to sink node and data aggregation tree is created by using accumulated pheromone. In [14] provides two different tree structures LPT and E-Span to facilitate aggregation of data in wireless sensor networks. In LPT, nodes having more remaining energy are chosen as aggregation parents. The tree is restructured when one node has no long function or when a broken link is identified. E-Span is an aware energy-spanning tree algorithm in which source node with maximal remaining energy is selected as root. Other source nodes select their corresponding parents from their neighbors on the basis of such information as

remaining energy and distance to root. In [15] an efficient energy-spanning tree is used to aggregate data in wireless sensor network for making which two parameters are used: energy and distance [15] uses route energy average to balance parameters energy and distance while previously provided algorithms have selected only one of these parameters as the main one and gave sound priority to the other. In [16] unlike common data aggregation methods, ESPDA avoids transmitting redundant data to head-clusters from sensor nodes in order to remove redundancy for improving application of efficient energy and bandwidth in sensor nodes. In [17] presents a scheme of efficient and highly accurate energy to aggregate data securely. The main idea of this is to aggregate data carefully without disclosing or reading secret information of sensors and posing considerable overhead in energy-limited sensors.

In [18] aggregation of data in wireless sensor networks is raised to balance latency and communication cost. In [19] spanning tree-based algorithms are provided to create high convergence between data aggregation and efficient energy and low latency in wireless sensor networks. Initially [19] provides two algorithms for making DAC tree. The first algorithm is the kind of minimum spanning tree, and the second of individual source shortest path spanning tree. Both of them are used as combined (COM) algorithm stimulator generally based on MST and SPT.

## 3. Preliminaries

Before presenting the algorithm, it is necessary to offer some primary definitions and concepts as follows.

### 3.1 Connected dominating set

Dominating set  $S$  of graph  $G = (V, E)$  is a host subset ( $s \subseteq v$ ) so that each host ( $v \subseteq V$ ) exists in set  $S$  or is adjacent to a host from  $S$ . In dominating set  $S$ , each host is referred to as dominator host, otherwise, as dominated host. A minimum dominating set is one with minimum cardinality. A connected dominating set  $S$  from graph  $G$  is one connected to each other. A minimum connected dominating set is a CDS with minimum cardinality.

There is no fixed, predefined infrastructure in wireless networks, and a virtual backbone can be formed by using CDS due to the lack of physical backbone. By definition, CDS is a subset of network nodes each of either belongs to CDS or is adjacent to at least one CDS node. This structure can be used to create a virtual backbone for routing and disseminating packets because it is connected. A minimum connected dominating set create a virtual backbone in graph, and reduces routing overhead considerably [20, 21].

### 3.2 Calculating sensors and CDS lifetimes

Let  $n$  be the number of sensors;  $B_i$  be initial energy of sensors;  $x_{i,j}$  be the number of bits being routed from sensor  $i$  to sensor  $j$ ;  $x_{i,0}$  be the number of bits being routed from sensor  $i$  to base station;  $t_{i,j}$  be sensor  $i$  communication cost of transmitting one bit to sensor  $j$ ; and  $r_{i,j}$  be sensor  $i$  communication cost for receiving one bit from sensor  $j$  [22]. In data aggregation process, each sensor receives data from one or more sensors, but sends data only to one sensor. Total sensor  $i$  consumed energy amount for transmitting one event ( $\xi_i$ ) is calculated as follows:

$$\xi_i = \sum_j X_{j,i} r_{j,i} + \sum_i X_{i,j} t_{i,j} \quad (1)$$

And sensor  $i$  remaining energy amount is calculated as follows:

$$B_i(t) = B_i(t - 1) - \xi_i \quad (2)$$

$B_i(t)$  And  $B_i(t - 1)$  are current and previous remaining energies, respectively.

Also, sensor  $i$  lifetime is calculated as follows:

$$\alpha_i = B_i / \xi_i \quad (3)$$

The average of CDS lifetime is calculated as follows:

$$\bar{\alpha}_{CDS} = \alpha_{CDS} / m \quad (4)$$

Where  $\bar{\alpha}_{CDS}$  equal to CDS lifetime average and  $\alpha_{CDS}$  equal to Total lifetime of sensors constituting CDS which is calculated as follows:

$$\alpha_{CDS} = \sum_{i=1}^m \alpha_i \quad (5)$$

In (4) and (5) equations  $m$  is the number of sensors exist in CDS.

### 3.3 Learning automata

A learning automata (LA) [23-25] is an abstract model capable of doing finite actions. Each selected action is evaluated by a probable environment, the result that is delivered to automata in the form of a positive or negative signal. Learning automata use this response to select their next action. Ultimate goal is for automatas to select the best of their actions. The best action is one maximizing the likelihood of receiving rewards from environment.

The Probable environment can be expressed mathematically by triple  $E = \{\alpha, \beta, \square\}$  where  $\alpha = \{\alpha_1, \alpha_2, \dots, \alpha_r\}$  is the set of environment inputs and  $\beta = \{\beta_1, \beta_2, \dots, \beta_r\}$  is each action's being penalized. Fig. 1 shows the relationship between learning automata and environment.

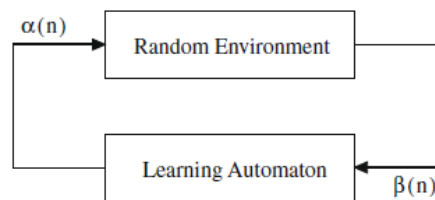


Fig. 1 the relationship between learning automata and environment Given the values of  $\beta$ , three different models are defined for probable environments. Whenever  $\beta$  is a two-members set of  $[0, 1]$ , the environment is of type  $p$ , that is, values of 0 and 1 is selected as environment outputs. In this case,  $\beta_1 = 1$  means "being penalized" and  $\beta_2 = 0$  means "being rewarded".

If  $\beta(n)$  is a value bounded to  $[0, 1]$ , the model is of type  $q$ ; and if  $\beta(n)$  is a stochastic variable within  $[0, 1]$ , the environment is of type  $S$ .  $C_i$  represents the probability that action  $\alpha_i$  receives an undesirable response from environment. The values of  $C_i$  do not change in static environments with changing the time in non- static ones [26].

Learning automatas are divided into two groups: (a) those with fixed structured, and (b) those with variable structured. In this paper, we make use of the variable structured. For learning Automatas with fixed structures, probabilities of automata actions are fixed while, for learning automatas with variable structures, they are updated with each turn of iteration. Learning automatas with variable structures can be denoted by triple  $\{\alpha, \beta, P, T\}$  where  $\alpha = \{\alpha_1, \alpha_2, \dots, \alpha_r\}$  is an automata's actions set;  $\beta = \{\beta_1, \beta_2, \dots, \beta_r\}$  is its inputs;  $P = \{P_1, P_2, \dots, P_r\}$  is probability vector of each automata's action; and  $\tau = p(x + 1) = T[\alpha(x), \beta(x), \rho(x)]$  is learning algorithm. Automatas choose their actions randomly on the basis of probability vector  $P_i$  and exercise. It is on the environments that they get a response. If the actions selected by Learning automate is action  $\alpha_i$ , then, automata updates its action probabilities to Eq. 6 in the case of receiving desirable response from environment while it does this according to Eq. 7 in the case of receiving undesirable one.

$$\begin{cases} P_i(n + 1) = P_i(n) + a[1 - P_i(n)] \\ P_j(n + 1) = (1 - a)P_j(n) \quad \forall j, j \neq i \end{cases} \quad (6)$$

$$\begin{cases} P_i(n + 1) = (1 - b)P_i(n) \\ P_j(n + 1) = \left(\frac{b}{r - 1}\right) + (1 - b)P_j(n) \quad \forall j, j \neq i \end{cases} \quad (7)$$

Where  $r$  is the number of automata's actions and  $b$  is penalty parameter. There following algorithms can be available on the basis of different values considered for parameters  $a$  and  $b$  of learning:

1) If  $a = b$ , linear reward-penalty (LR - P) scheme is obtained.

- 2) If the value of  $b$  is many times smaller than that of  $a$ , resulting learning method is called liner reward epsilon scheme ( $LR_{\epsilon P}$ ).
- 3) If  $b = 0$ , algorithm is called linear reward inaction ( $LR_I$ )

### 3.3.1 Distributed learning automata (DLA)

A distributed learning of automata (DLA) [27, 28] is a network of LAs cooperating to solve a particular problem. Within this network of cooperating automata, only one automata is active at a time. In DLA the number of actions each automata is able to do is equal to the number of automatas connects to that one. When an automata selects an action in the network, other automata connected to it is activated. In other words, Choosing an action by an automata in this networks corresponds to activation of another automata there. The model considered for DLA network is graph each vertex of which is an automata, as shown in fig. 2 In this graph, presence of edge ( $LA_i, LA_j$ ) means that choosing the action  $\alpha_j^i$  by  $LA_i$  activate  $LA_j$ . The number of actions  $LA_k$  can select is denoted as  $P^k = \{P_1^k, P_2^k, \dots, P_{r_k}^k\}$ . within this set,  $P_m^k$  represents probability related to action  $\alpha_m^k$ . Selecting the action  $\alpha_m^k$  by  $LA_k$  activates  $LA_m$ .  $r_k$  Shows the number of actions  $LA_k$  is able to do.

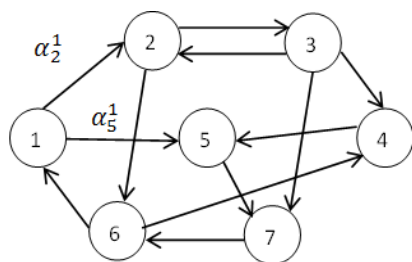


Fig. 2 Network of distributed learning automatas

## 4. Forming virtual backbone based on distributed learning automata

Suppose that wireless sensor network includes a group of wireless hosts having transmission range  $r$  and linking, directly or indirectly, to each other. Here, suppose that topological graph corresponds with unit graph where host 1 ( $H_1$ ) corresponds with vertex 1 ( $V_1$ ). Any two hosts connected to each other are said to be neighbors having mutual communication. Therefore, it is assumed that network graph is a undirected graph. Each host has a unique argument and is required to know its neighbors. In this section, an algorithm based on distributed learning automata is provided for data aggregating in wireless sensor networks, focusing on finding an almost optimal solution for problem of data aggregation in network graph. In this approach, each host (e.g.  $H_i$ ) is equipped with learning automata (e.g.  $A_i$ ). A network of

learning of automata is denoted by binary  $\langle A, \alpha \rangle$  where  $A = \{A_1, A_2, \dots, A_n\}$  indicates set of learning automatas corresponding to set of vertexes (hosts) and  $\alpha = \{\alpha_1, \alpha_2, \dots, \alpha_n\}$  represents action set. And also,  $\alpha_1 = \{\alpha_{11}, \alpha_{12}, \dots, \alpha_{1n}\}$  represents an action set which can be run by learning automata  $A_1$ . Here, we use learning automata with variable actions the number of which depends on the number of adjacent vertexes (neighbors) of respective learning automata.

### 4.1 Action set formation method

In the algorithm provided for forming action set related to learning automata  $A_i$ , initially, its host (e.g.  $H_i$ ) sends a message locally to its neighbors one step apart locally. Hosts located in transmission range from sending host respond to it upon receiving the message and send back their action sets to primary sending host which creates its own action set on the basis of responses received from neighbors. Therefore, each host  $H_i$  the message of which was responded adds action  $\alpha_i$  to action set of learning automata  $A_i$ . In fact, when  $H_i$  sends a message to  $H_j$  which sends back its response to  $H_i$ , the learning automata corresponding to  $H_i$  adds action  $\alpha_j$  (selection of vertex  $j$  ( $V_j$ )) to action set of its own corresponding automata (namely,  $A_i$ ). Selection of action  $\alpha_j$  corresponding to  $H_i$  as dominator is performed by learning automata  $A_i$ . So the size of each learning automata action set depends on the order of respective host, assuming that hosts have been distributed in network uniformly. A problem with above defined action set, in which the number of actions is fixed and does not change with time, may result in frequent selection of a host, with virtual backbone including redundant loops and dominators. Therefore, fixed action sets decrease convergence speed of algorithm and enlarge the size of virtual backbone. To overcome these shortcomings, we suggest learning automata with variable actions and present following rules for pruning action set of learning automata.

**Rule1.** To avoid choosing the same dominators (by different hosts), each activated learning automaton is allowed to prune its action-set by disabling the actions corresponding to the dominator hosts selected earlier. This rule increases the convergence speed, and consequently, decreases the running time of the proposed algorithm.

**Rule2.** To avoid the loops and the redundant dominator hosts by no more (dominate) hosts can be spanned; the proposed algorithm prunes the action-set as follows.

As mentioned earlier, when host  $H_i$  is going to form the action-set of its automaton, it receives some messages from its neighboring hosts which include the action-set information of these hosts. Depending on received information, activated automaton  $A_i$  updates its action-set by disabling the actions corresponding to the hosts whose one-hop neighbors

all have been spanned (or added to the dominate set) before (see Fig. 4a–f), if any. This rule reduces the dominator set (backbone) size, decreases the running time and improves the convergence rate of algorithm.

## 4.2 Algorithm description

As mentioned above, we consider our network as a unit disk graph that is, sending radiuses of all graph nodes are equal. Also, we assume the nodes are distributed in network graph randomly. Each of sensor nodes possesses some amount of energy being approximately the same for all nodes at first. Over time, the level of nodes energy changes. In this network, sensor nodes have some information about themselves and their neighboring nodes, including their energy level at a given time, which is updated periodically. As stated earlier, nodes participating in CDS are referred to as dominators and the rest are dominates. We create CDS in accordance with energy levels of network nodes. In fact, we use nodes with higher levels of energy to create CDS. Our aim is actually to make created route more permanent. In other words, we want to increase created CDS lifetime. Each node of network is equipped with learning automata; therefore, we have a network of learning automata each of which has a selected action set. The numbers of each automaton's operations are equal with the number of nodes neighboring the node corresponding to targeted automata that can select only one action from its action set at a moment over time. Given the decrease in nodes energies, we calculate their energy periodically. The amount of energy consumed for sending or receiving a message differs. The amount of energy usually consumed to send a message is much more than that consumed to

receive it. These nodes not located on the route of made CDS go into idle (or sleep) state in which their consumption energy is near zero, thus, energy consumption is associated with nodes located on the route of made CDS. A CDS is made any time the algorithm is iterated. Initially, we define an energy threshold for nodes existing in network, which is the least amount of energy needed by each network node which is for network permanence.

The level of energy of each of nodes located on made CDS route should not be less than this amount of energy defined as threshold level. If so, learning automata of nodes located on the route are penalized, if not, are rewarded. So that the probability of selecting these nodes to make future CDS routes increases. The process of CDS-making continues until made CDS converges toward an optimal response. The pseudo code of algorithm is presented below (Fig. 3). Here,  $m$  and  $k$  represent the number of nodes constituting CDS and the number of steps of making CDS, respectively.

$CDS_k$  Is dynamic threshold at  $k$ th step;  $CDS_k$  is calculated with Eq. 8. CDS is the selected connected dominating set;  $W_{cds}$  is the weight of made CDS and calculated with Eq. 9;  $V_{A_i}$  is the vertex corresponding to learning automata  $A_i$ ;  $N_{V_{A_i}}$  represents neighbors adjacent to vertex  $i$  ( $V_i$ ); and  $r$  is the number of CDSs made until step  $k > 1$ .

$$CDS_k = \frac{1}{r} \sum_{i=1}^r W_{cds} \quad (8)$$

$$W_{cds} = 1/\alpha_{cds} \quad (9)$$

---

### Algorithm for forming a CDS with maximum of lifetime

---

- 1: Input: Graph  $\langle V, E, W \rangle$ ,  $P_{cds}$ , Iteration\_max
- 2: Output: Optimal nearest Data aggregation
- 3: Assumptions
- 4: let CDS denotes the selected connected dominated set
- 5: Begin algorithm
- 6:  $k \leftarrow 0$ ,  $CDS_k \leftarrow 0$
- 7: Repeat
- 8:  $CDS \leftarrow \emptyset$ ,  $W_{cds} \leftarrow 0$ ,  $Dominator\_set \leftarrow \emptyset$ ,  $Dominatee\_set \leftarrow \emptyset$
- 9: the automaton corresponding to sink node is selected, denoted as  $A_i$  and activated,  $m = 1$
- 10:  $Dominator\_set \leftarrow Dominator\_set + V_{A_i}$ ,  $Dominatee\_set \leftarrow Dominatee\_set + V_{A_i} + \{N_{V_{A_i}}\}$ ,  
 $CDS \leftarrow CDS + V_{A_i}$ ,  $\alpha_{cds} \leftarrow \alpha_{cds} + \alpha_{V_{A_i}}$
- 11: Repeat
- 12: If ( $|Dominatee\_set| \neq Network\ size \ \&\& \ A_i \ has \ no \ possible \ actions$ ) Then
- 13: Path induced by active automata is traced back for finding an automata with available actions
- 14: the found learning automata is denoted as  $A_i$
- 15: End If
- 16: Each automata prunes its action set
- 17: automaton  $A_i$  chosen one of its actions
- 18:  $Dominator\_set \leftarrow Dominator\_set + V_{A_i}$ ,  $Dominatee\_set, Dominatee\_set + V_{A_i} + \{N_{V_{A_i}}\}$ ,  $CDS \leftarrow CDS + V_{A_i}$ ,  $\alpha_{cds} \leftarrow \alpha_{cds} + \alpha_{V_{A_i}}$ ,  $m = m + 1$
- 19: automaton  $A_j$  is active
- 20: set  $A_i$  to  $A_j$
- 21: until ( $|Dominatee\_set| \neq Network\ size \ \&\& \ A_i \ has \ no \ possible \ actions$ )

```

22:  $\bar{\alpha}_{cds} = \alpha_{css}/m$ ,  $W_{cds} = 1/\bar{\alpha}_{cds}$ 
23: Data aggregation
24: compute the average weight of CDS and denote it  $W_{cds}$ 
25: If ( $W_{cds} < CDS_{k-1}$ ) Then
26: Reward the selected actions of the activated automata along the CDS
27: Else
28: Penalize selected actions of the activated automata along the CDS
29: End If
30:  $CDS_k \leftarrow [(k-1)CDS_{k-1} + W_{cds}] / k$ 
31:  $k \leftarrow k + 1$ 
32: Enable all the disabled actions
33: until ( $k \geq Iteration\_Max$  OR probability of finding dominator set  $> Pc_{ds}$ )
34: End Algorithm
    
```

Fig. 3 The pseudo code of proposed algorithm

Fig. 4a-e illustrates the steps of making backbone in the first implementation of the algorithm. In fig. 4b, node 3 is considered as sink node and the first dominator. In this case, nodes 3, 1, 4 and 5 are added to dominate set. In fig. 4c, node 4 is selected as next dominator and node 6 is added to dominate set. In fig. 4d, node 6 is selected as next dominator and nodes 2, 7, 8 and 10 are added to dominate set. In fig. 4e node 2 is selected as next dominator and node 0 is added to dominate set. Up to this step, we have dominator set of {2,6,4,3} and dominate set of {0,10,8,7,2,6,5,4,1,3}. Entire

network is not still covered because the size of dominate set is smaller than network's; therefore, the algorithm performs backtracking and selects node 10 as next dominator, hence nodes 9 and 11 added to dominate set (fig. 4f). Now, we have dominator and dominate sets of {10,2,6,4,3} and of {11,9,0,10,8,7,2,6,5,4,1,3}, and entire network is covered. Fig. 4g and 4h were obtained in the second and third implementation of algorithm, respectively, that node 5 is considered as sink node and the first dominator.

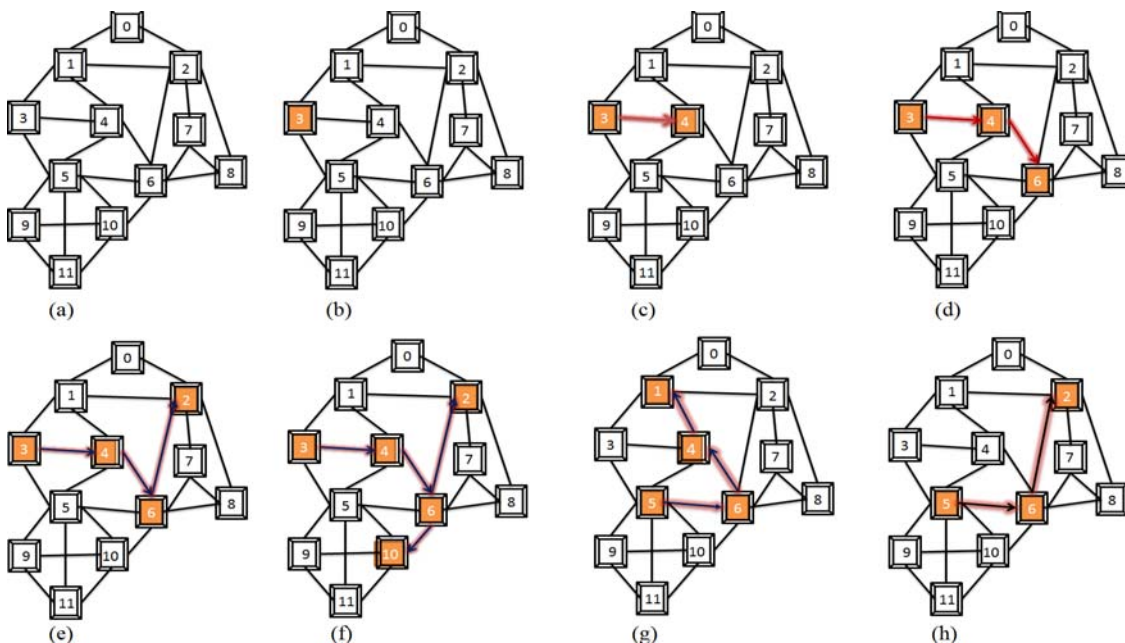


Fig. 4 The step-by-step backbone formation process

## 5. Simulation results

In this paper, NS2 software was used to simulate wireless sensor network. Simulation was performed in a square area of  $100 \times 100 m$  with 50, 100 and 150 nodes distributed uniformly in the environment. We assumed learning rate is 0.2 and initial energy is  $500mj$  for each node. Also, we assumed each node consumes  $6mj$  and  $3mj$  units of energy to send and receive any kinds of packets, respectively. For this simulation, the

threshold of CDS process and max iteration were set at 0.9 and 200, respectively. In here, for assessing our algorithm (CDS-LT), we compare our algorithm with proposed algorithms in [14] and [19]. In [14] Lee and Wong have proposed two different tree structures LPT and E-Span to facilitate aggregation of data in wireless sensor networks. In LPT, nodes having more remaining energy are chosen as aggregation parents. The tree is restructured when one node has no long function or when a broken link is identified.

In [19] Upadhyayula and Gupta have proposed spanning tree-based algorithms are provided to create

high convergence between data aggregation and efficient energy and low latency in wireless sensor networks. Initially [19] provides two algorithms for making DAC tree. One of the algorithms is the kind of individual source shortest path spanning tree. In here, we evaluate our simulation with respect to made CDS lifetime by expanding transmission range and increasing the number of nodes. We assume transmission range changes from 10m to 20m. As it is

show in fig. 5, the CDS lifetime decreases when the transmission range increase. And CDS lifetime decreases by increasing the number of nodes. Also with comparing our algorithm (CDS-LT) with proposed algorithms in [14] and [19] will determine how much our method performs well.

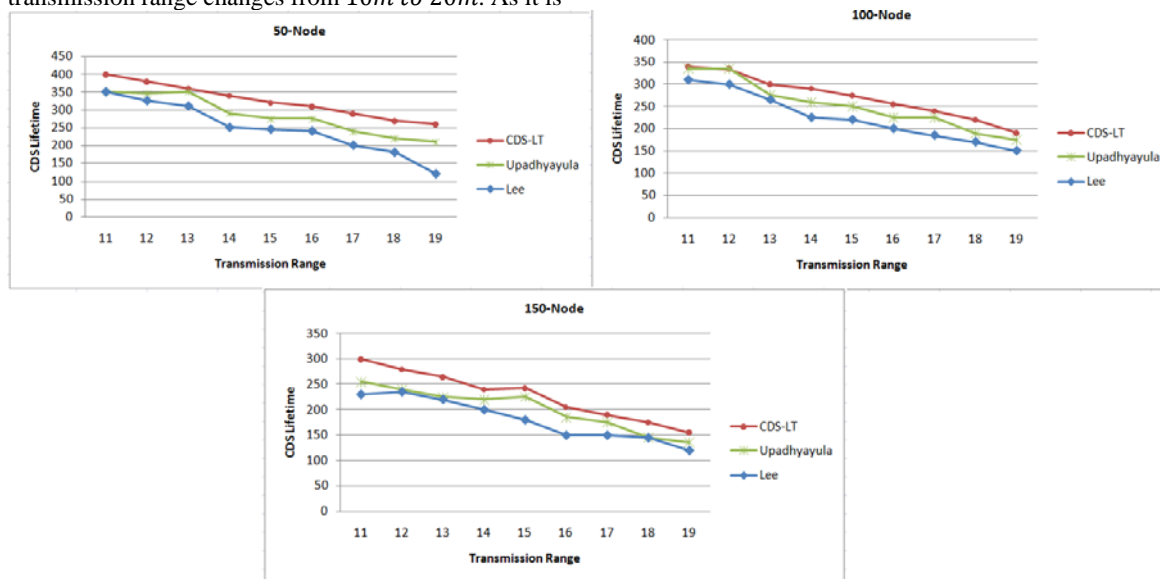


Fig. 5 Comparison CDS life time for CDS-LT algorithm

## 6. Conclusion

This paper is provided on heuristic algorithm based on distributed learning automata to solve problems of data aggregation in stochastic graphs. Given that data aggregating by creating backbones and making CDSs in networks lowers the ratio of responding hosts to the hosts existing in virtual backbones, we used this idea in our algorithm and tried to find CDSs with the

longest lifetime considering such parameters as lifetime, remaining and consumption energies of sensors in order to have an optimal data aggregation. Simulation results showed that lifetime of made CDSs decreased as the number of nodes increased and transmission range expanded. Also, we compared our algorithm with proposed algorithms in [14] and [19], as shown above our algorithm always outperforms the others in terms of the life time.

## References

- [1] P. Gupta, and P. R. Kumar, "The capacity of wireless networks", IEEE Transaction on Information Theory, Vol. 46, No. 2, 2000, pp. 388–404.
- [2] Y. Z. Chen and A. L. Liestman, "Approximating minimum size weakly connected dominating sets for clustering mobile ad hoc networks", in: Proceedings of the Third ACM International Symposium on Mobile Ad Hoc Networking and Computing, 2002, pp. 157–164.
- [3] J. Wu, F. Dai, M. Gao and I. Stojmenovic, "On calculating power-aware connected dominating sets for efficient routing in ad hoc wireless networks", Journal of Communications and Networks, Vol. 4, No. 1, 2002.
- [4] Y. P. Chen, and A. L. Liestman, "Maintaining weakly-connected dominating sets for clustering ad hoc networks, Ad Hoc Networks, Vol. 3, 2005, pp. 629–642.
- [5] B. Han, and W. Jia, "Clustering wireless ad hoc networks with weakly connected dominating set", Journal of Parallel and Distributed Computing, Vol. 67, 2007, pp. 727–737.
- [6] B.N. Clark, C.J. Colbourn and D.S. Johnson, "Unit disk graphs", Discrete Mathematics, 86, 1990, pp. 165–177.
- [7] M.V. Marathe, H. Breu, H. B. Hunt III, S. S. Ravi and D. J. Rosenkrantz, "Simple heuristics for unit disk graphs", Networks, Vol. 25, 1995, pp. 59–68.
- [8] R. Shah, and J. Rabaey, "Energy Aware Routing for Low Energy Ad Hoc Sensor Networks", Communication /Computation Piconodes for Sensor Networks, 2002.
- [9] M. Esnaashari and M. R. Meybodi, "Data Aggregation in wireless Sensor Networks using Learning Automata", Wireless Netw, vol.16, 2010, pp. 687–699.
- [10] A. Karaki, and R. Ui-Mustafa, et al, "Data aggregation and routing in Wireless Sensor Networks: Optimal and heuristic algorithms", Computer Communications, 2009, pp. 945–960.

- [11] S. ozdemir and Y. Xiao. "Secure data aggregation in Wireless Sensor Networks:A comprehensive overview" *computer Networks*, 2009, pp. 2022-2037.
- [12] S. Soro and W. B. Heinzelman, "Prolonging the Lifetime of Wireless Sensor Networks via Unequal Clustering", *IEEE*, pp. 2005.
- [13] W. H. Liao, Y. Kao, and et al. "Data aggregation in wireless sensor networks using ant colony algorithm." *Network and Computer Applications*, 2008, 387-401.
- [14] W. M. Lee and V. W. S. Wong, "E-Span and LPT for data aggregation in wireless sensor networks", *Computer Communications*, 2006, pp. 2506-2520.
- [15] Z. Esjkandari, M. H. Yaghmaee, et al, "Energy Efficient Spanning Tree for Data Aggregation in wireless sensor networks", *IEEE*, 2009.
- [16] H. Cam, S. Ozdemir, and et al, "Energy-efficient secure Pattern based data aggregation for wireless sensor networks", *Computer Communications*, 2006, pp. 446-455.
- [17] H. Li, K. Lin, and et al, "Energy-efficient and high-accuracy secure data aggregation in wireless sensor networks", *Computer Communications*, 2010.
- [18] P. Korteweg, A. Marchetti-Spaccamela, and et al, "Data aggregation in sensor networks:Balancing communication and delay costs", *Theoretical Computer Science*, 2009, pp. 1346-1354.
- [19] S. Upadhyayula and S. K. S. Gupta, "Spanning tree based algorithms for low latency and energy efficient data aggregation enhanced convergecast(DAC) in wireless sensor networks", *Ad Hoc Networks* vol. Vol. 5, 2007, pp. 626-648.
- [20] K. M. Alzoubi, P. J. Wan and O. Frieder, "Maximal independent set, weakly connected dominating set, and induced spanners for mobile ad hoc networks", *International Journal of Foundations of Computer Science*, Vol. 14, No. 2, 2003, 287–303.
- [21] S. Basagni, M. Mastrogiovanni, C. Petrioli, "A performance comparison of protocols for clustering and backbone formation in large scale adhoc network", in: *Proceedings of the First IEEE International Conference on Mobile Ad Hoc and Sensor Systems*, 2004, pp. 70–79.
- [22] J. C. Dagher, M. W. Marcellin and M. A. Neifeld, "A Theory for Maximizing the Lifetime of Sensor Networks", *IEEE Transaction on Communications*, Vol. 55, No. 2, 2007.
- [23] K. S. Narendra and K. S. Thathachar, "Learning Automata: An Introduction", *Prentice-Hall*, New York, 1989.
- [24] M. A. L. Thathachar and P. S. Sastry, "A hierarchical system of learning automata that can learn the globally optimal path", *Information Science*, Vol. 42, 1997, pp. 743–766.
- [25] M. A. L. Thathachar and B. R. Harita, "Learning automata with changing number of actions", *IEEE Transactions on Systems, Man, and Cybernetics SMG*, Vol. 17, 1987, pp. 1095–1100.
- [26] S. Lakshmiarahan and M. A .L. Thathachar, "Bounds on the convergence probabilities of learning automata", *IEEE Transactions on Systems, Man, and Cybernetics SMC*, Vol. 6, 1976, pp. 756–763.
- [27] K. S. Narendra and M. A .L. Thathachar, "On the behavior of a learning automaton in a changing environment with application to telephone traffic routing", *IEEE Transactions on Systems, Man, and Cybernetics SMC*, Vol. 10, No. 5, 1980, pp. 262–269.
- [28] H. Beigy and M. R. Meybodi, "Utilizing distributed learning automata to solve stochastic shortest path problems", *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 14, 2006, pp. 591–615.



**Chamran Asgari** received the B.S. and M.S. degrees in Computer Engineering in 2009 and 2012, respectively, Arak, Iran. His research interests include data aggregation algorithms in wireless sensor networks, learning systems, parallel algorithms, virtual Backbone formation and minimum spanning tree.

# A 100 mA Low Voltage Linear Regulators for Systems on Chip Applications Using 0.18 $\mu\text{m}$ CMOS Technology

Krit Salah-ddine<sup>1</sup>, Zared Kamal<sup>2</sup>, Qjidaa Hassan<sup>3</sup> and Zouak Mohcine<sup>4</sup>

<sup>1</sup>University Ibn Zohr Agadir Polydisciplinary Faculty of Ouarzazate Morocco.

<sup>2</sup> University Sidi Mohamed Ben Abdellah, Technical and Sciences Faculty,  
Laboratory of Signals,  
Systems and Components, Morocco.

<sup>3</sup> University Sidi Mohamed Ben Abdellah, Faculty of Sciences Dhar El Mehraz Fez  
Laboratory of Electronic Signals –Systems and Informatics (LESSI), Morocco.

<sup>4</sup> University Sidi Mohamed Ben Abdellah, Technical and Sciences Faculty,  
laboratory of Signals, Systems and Components, Morocco.

## Abstract

A novel design for a low dropout (LDO) voltage regulator is presented and dedicated to power many sections of a typical cellular handset. However, these baseband, RF, and audio sections have different requirements that influence which LDO is most appropriate. After discussion of the specific requirements, different LDOs are recommended. Also, some LDO design techniques are briefly discussed to demonstrate how an LDO may be optimized for a specific level of performance.

Cellular phone designs require linear regulators with low-dropout, low-noise, high PSRR, low quiescent current ( $I_q$ ), and low-cost. They need to deliver a stable output and use small-value output capacitors. Ideally, one device would have all these characteristics and one low-dropout linear regulator (LDO) could be used anywhere in the phone without worry. But in practice, the various cell phone blocks are best powered by LDOs with different performance characteristics. This paper provides a new design methodology to choosing the right LDO to power each cell phone and especially for the Voltage Phase-Locked loops (VPLLs) blocks. Fabricated in a 0.18  $\mu\text{m}$  CMOS process, the measured results show the adopted topology achieves a better phase noise than the conventional saturation current source. and the spread of the current limitation (without matching) is 100mA, the VPLLs system demonstrates a phase noise of 782  $\text{nv}/\sqrt{\text{Hz}}$  at 100-kHz, and 33  $\text{nv}/\sqrt{\text{Hz}}$  at 1 MHz, while quiescent current 33  $\mu\text{A}$  from a 2.6 V supply voltage.

## Key words:

LDO, PSRR, low noise, cell phone, handset, RF, baseband, audio, GSM

## 1. Introduction

Low dropout regulators (LDOs) are widely used and implemented in most circuit applications to provide

regulated power supplies. The increasing demand of performance is especially apparent in mobile battery-operated products, such as cellular phones, pagers, camera recorders, and laptops [1-7]. For these products, very high PSRR, low noise regulators are needed. Moreover such high-performance regulators have to be designed in standard low-cost CMOS process, which makes them difficult to realize. For PSRR point of view, as depicted in [2], this kind of regulator requires a first-stage amplifier with a large gain-bandwidth product (Product of its dc-gain and cut-off frequency, which is typically 10 MHz). This first-stage amplifier performance can be achieved either by a large dc-gain, or by a high cut-off frequency. Compared with switching regulators, LDOs are less expensive, smaller in size and easier to be used. Moreover, the noise of output voltage is lower and the response to input voltage transient and output load transient is faster. These advantages make LDOs suitable for battery-powered equipments, communication systems, portable systems, and post regulators of switching regulators. Among possible process technologies, CMOS technology



is very attractive for LDO circuit implementation because of its low cost, low power consumption and potential for future system-on-chip integration.

In this paper, a CMOS LDO using new scheme that can maintain system stability with a maximum load current limitation =100 mA is proposed figure1shows the regulator integrated in the system on chip and figure2 shows the schematic of the proposed regulator with power transistor who provide 100mA.

Moreover, it can provide stable output under all load conditions with a value of load capacitor equal 2.2 uF. The ESR of the load capacitor can range from zero to some finite value equal 100 mohm.

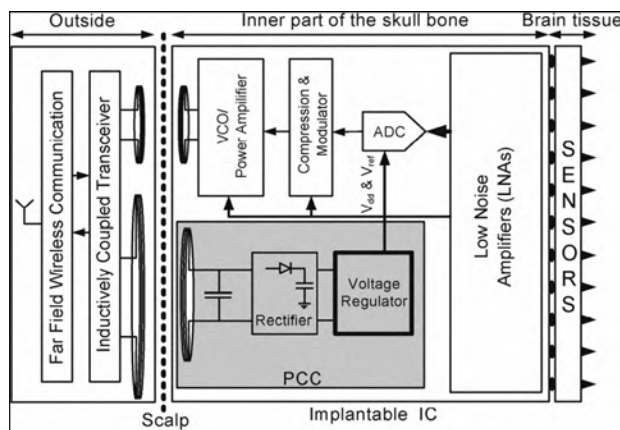


Fig.1 regulator integrated in the system on chip

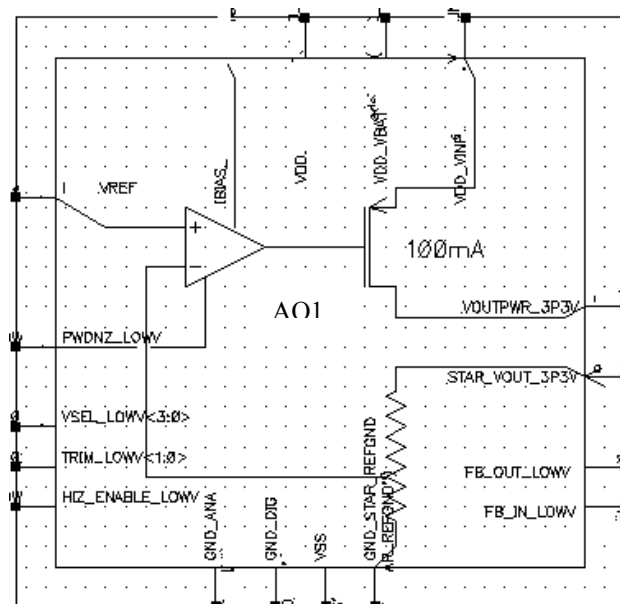


Fig.2 schematic of the proposed regulator with power transistor who provide 100mA.

## 2. Ldo Characterizations

LDO design involves three primary aspects, namely, regulating performance, current efficiency, and operating voltage [1]. These design aspects are explicitly stated in the following design specifications: 1) dropout, 2) line regulation, 3) load regulation, 4) temperature dependence, 5) transient output voltage variation as a result of load current steps, 6) quiescent current, and 7) power supply rejection ratio [1]. Table 1 lists the target specifications for the three LDO architectures. Each of these specifications is discussed in this section.

### i. Dropout

Dropout is the minimum input to output voltage difference at which the LDO ceases to regulate, determined by unacceptable decrease in output voltage. If the maximum current and minimum dropout conditions are not satisfied within the error margin, the LDO is then not performing its regulating function properly. The pass device must be large enough to guarantee the minimum dropout (source-to-drain or input-to-output voltage difference) while providing the maximum load current. Dropout is simulated by performing a dc sweep on the input voltage and plotting the output voltage. Dropout is the voltage difference between the input voltage and the output voltage at the point where the input voltage is minimum (2.6 V in this case). This point is estimated to be at the minimum input voltage at which the LDO is allowed to operate.

### ii. Line Regulation

Line regulation is the output voltage change as a result of a specific change in input voltage at a specific load current. Line regulation is simulated by performing a dc sweep on the input voltage and plotting the output voltage, is measured at both the maximum and minimum load currents.

### iii. Load Regulation

Load regulation is the ratio of the change in output voltage to the change in load current, which is the regulator output resistance,  $R_{LDO}$ . Load regulation is simulated by performing a dc sweep on the load current and plotting the output voltage. Then the load regulation of the regulator is  $R_{oLm} = 0,001mV$

### iv. Temperature Dependence

Temperature dependence is the change in the output voltage due to a change in temperature. LDO temperature dependence is a function of the temperature dependence of the reference voltage and the offset voltage of the error amplifier. The temperature coefficient of a CMOS band-gap voltage reference can be as low as 15 ppm/°C over a temperature range from -40°C to 125°C [2]. Therefore, the temperature dependence of the band-gap reference voltage is ignored in this simulation. LDO temperature

dependence is simulated with a dc sweep on the temperature and plotting the output voltage.

#### v. Transient Output Voltage Variation

Transient output voltage variation is the output voltage change in response to transient load current variation and is a function of four parameters: system time response, output capacitance, maximum load current, and ESR of the output capacitor. Transient output voltage variation is simulated by applying a transient load current signal and plotting the transient output voltage, is measured for a certain  $A_t$  (rising and falling time of the transient load current signal).

#### vi. Quiescent Current

Quiescent current is the total current drawn from the voltage supply at zero load current. Quiescent current is simulated by performing a dc operating point and measuring the total current consumed by the EA and the feedback resistors.

#### vii. Power Supply Rejection Ratio

Power supply rejection ratio (PSRR) is the ratio of the change in output voltage to the change in input voltage supply. It is also defined as the ac voltage gain from the input node to the output node of the LDO regulator. PSRR is simulated by performing an ac sweep of the input voltage supply and plotting the ratio of the output voltage to the input voltage. PSRR is measured at the frequencies of interest in dB

## 2. Design of Current Limiting Circuit

### 2.1 Design Requirement of Current Limiting Circuit

A current limiting circuit used in LDO linear voltage regulator should at least meet the requirements as follows:

- When overcurrent hasn't taken place, the voltage regulator should regulate the output voltage  $V_{out}$  normally, and the current limiting circuit should have little effect on it.
- A current limiting circuit should first include output current detecting devices or block to detect if output current  $I_O$  has exceeded the maximum rated value.
- After the current limiting circuit starts up, it should cut off the negative feedback loop of the regulator. Then the regulator cannot regulate the output voltage any more.
- After foldback current limiting circuit starts up,  $I_O$  (Input Output) will decrease as  $V_{out}$  decreases. As the output is shorted,  $I_O$  will be limited to a value much less than the maximum rated value.

Besides, a good current limiting circuit should take some other factors into consideration, such as: low quiescent current and power consumption, few devices, low cost, and soon.

### 2.2 Design Principle of Current Limiting Circuit

The current limiting circuit presented in the paper is showed in Fig. 1. It comprises output current sampling circuit, constant current limiting circuit and foldback current limiting circuit. Signals VB1 and VB2 are generated by the self-biasing circuit of the error amplifier (We only give the second stage of the amplifier in Fig. 1). The potential of VB2 is constant, and  $VSG\_MP = VDD - VB1$  holds constant as well.

MN1 and MP2 make up of the second stage of the error amplifier. AO1 is its input as well as the output of the first stage and AMP\_OUT is its output as well as the output of the error amplifier. PW is pass element. PWDNZ is the enable control signal. When it is at high potential, MP1 is off and the circuit works normally.

## 3. Simulation and experimental Results of the proposed Voltage Regulator

the architecture of the proposed voltage regulator as shows in Figure 3 which reduces the total cost and facilitates the regulator implantation. The supply voltage denoted as  $V_{in}$  is provided by the rectifier output, and can be as low as 2.6V. the reference voltages  $V_{REF} = 0.75V$  as bandgap reference circuit [5] with dynamic start-up and turn-on time circuitry is used to generate the required reference voltages and currents the simulation results is shown in figure 4. The bandgap is supplied from the regulator output, which mitigates the need for high PSRR reference voltage generation.

the current limitation of the proposed regulator LDO was simulated and realized from the structure depicted in fig.3, using a  $2.2\mu F$  external capacitor on  $V_{OUT}$ . This regulator was designed to deliver 3.40V with a maximum load current of 100mA.

Figure 5 shows the simulation results of the output noise and PSRR outputs, when the  $V_{DD}$  voltage is rising and falling. This indicates, as explained previously, that when  $V_{DD}$  rises and is below 2.35V, the POR signal stays low, and forces  $V_{OUT}$  to follow  $V_{DD}$ . In these conditions, the total quiescent current for this circuit is below  $1\mu A$ . For higher values of  $V_{DD}$ , the LDO regulates the output voltage  $V_{OUT}$  to 2.40V. The maximum quiescent current is obtained for  $V_{DD} = 5.5V$ , and is equal to  $1.5\mu A$ .

The figure 6 shows that the the phase margin versus  $i_{load}$  who verify the stability of our architecture. Figure 7 schows the dc-ligne regulation of the regulator.

Then this ultra-low quiescent regulator including the POR was fabricated in a CMOS 0.18  $\mu m$  process. It has been optimized for quiescent current.

The external load capacitor was  $2.2\mu F$ , the output load current is less or equal than 1mA, and the input voltage  $V_{DD}$  is below 5.5V.

The measured total quiescent current was less than  $2\mu\text{A}$  for  $V_{DD}$  in the  $0\text{V}-5.5\text{V}$  range. These measured values are in good accordance with the simulated results above.

Table 1: Margin specifications

Parameters	Typic Value	Units
Vin	2.6	V
Ilimit	100	mA
Ligne R	5	mV
Load R	20	mV
PSRR	65	dB
DC gain	89	dB

Table 2: Performance comparison between recent works on SoC LDOs

	[14]	[15]	[16]	[17]	This work
Year	2003	2007	2007	2009	2011
Process	$0.6\mu\text{m}$	$0.35\mu\text{m}$	$0.35\mu\text{m}$	$0.35\mu\text{m}$	$0.18\mu\text{m}$
Vin	1.5V	3V	1.2-3.3 V	1.2-1.5 V	2.6 V
Vout	1.3 V	2.8 V	1V	1V	2.4 V
$I_Q$	$38\mu\text{A}$	$65\mu\text{A}$	$100\mu\text{A}$	$45\mu\text{A}$	$1.5\mu\text{A}$
$I_L^{\text{MAX}}$	100 mA	50 mA	100mA	50mA	100mA
$\Delta V_{\text{out}}$	100mV	< 90mV	50mV	70mV	20mV

#### 4. Performance Comparison

Table 2 provides comparison between the performance of the proposed LDO regulator and other published designs that are targeted for SoC power management.

Fig.3 Architecture of the proposed regulator

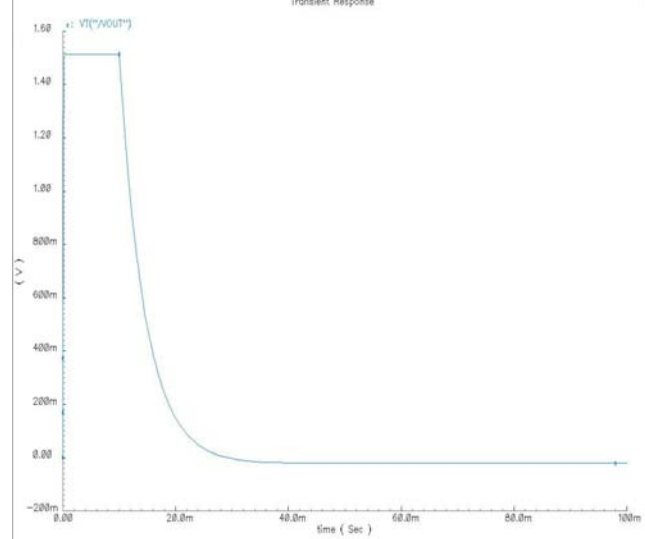
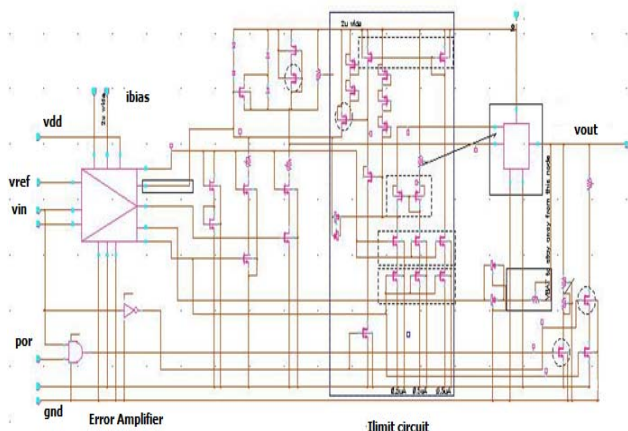
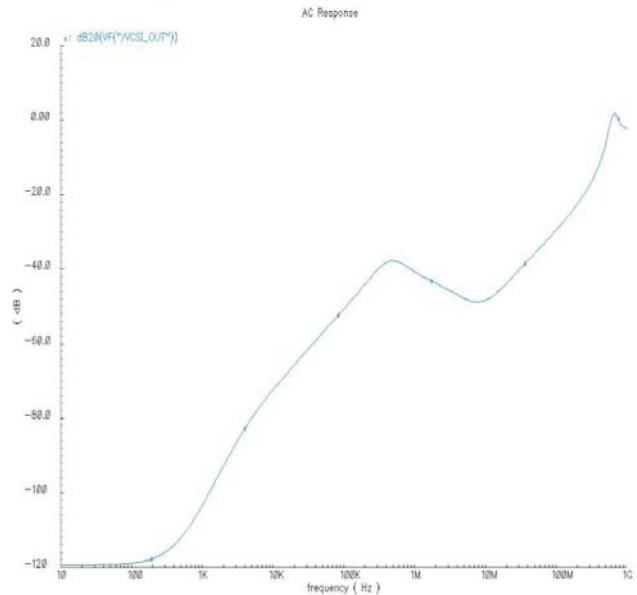


Fig.4 Startup and Turn-Off



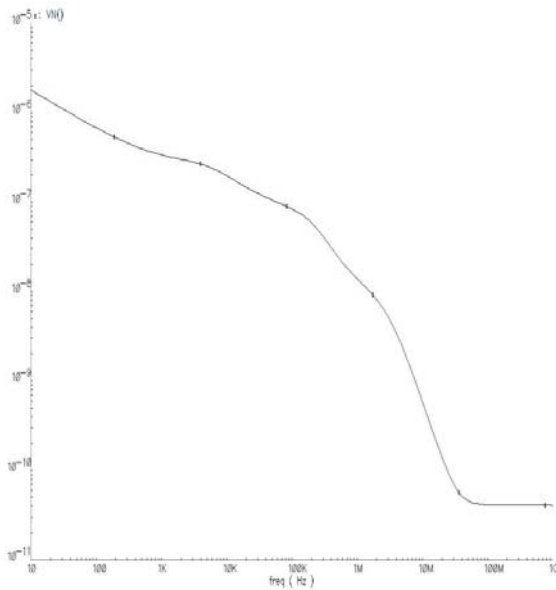


Fig.5 Output Noise and PSRR

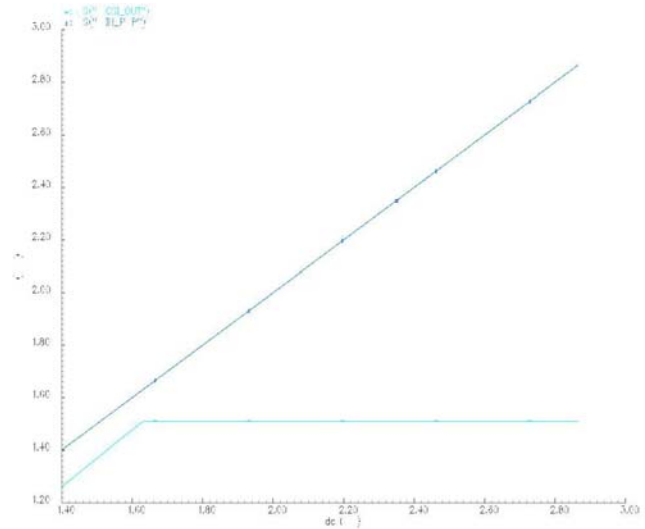


Fig.7 Dc ligne Regulation

## 5. Abreviations

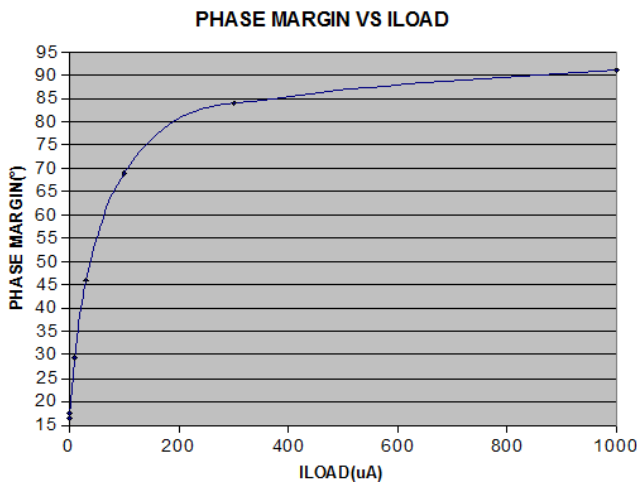


Fig.6 Phase margin versus iload

abbreviations	Sens
PSRR	Power Supply Rejection Ratio
VPLLs	Voltage Phase-Locked Loops
CMOS	Complementary Metal-Oxyde Semiconductor
LDO	Low Drop Out
ESR	Equivalent Series Resistor
DC	Direct Current
Iload	Current Load
VB	FeedBack
VDD	Voltage positive supply
VREF	Voltage Reference
VOUT	Output Voltage
PWDNZ	Power Down Zero: this input can be driven low (at a logic 0)
POR	Power On Reset
SoC	System on Chip
I <sub>Q</sub>	Quiescent Current

Table 3: Abbreviations of the words used in this proposed work.

## References

- [1] <http://www.ieice.org/eng/shiori/mokuji.html>
- [2] Ernst Lueder, "Liquid Crystal Displays", John Wiley & Sons, 2001
- [3] R. Harrison, P. T. Watkins, R. I. Kier, R. O. Lovejoy, D. J. Black, B. Greger, and F. Solzbacher "A Low-Power Integrated Circuits for a wireless 100-Electrode Neural Recording System," IEEE I. Solid-State Circuits, vol. 42, no. 1, pp. 123-133, Jan. 2007.
- [4] T. Ying, W.H. Ki, and M. Chan, "Area Efficient CMOS Integrated Charge ps", ISCAS 2002, IEEE International

- Symposium on Circuits and Systems, Vol. 3, pp. 831-834, 2002
- [5] G. Thiele and E. Bayer, "Current Mode Charge Pump, Topology, Model and Control", IEEE Power Electronics Specialists Conference, pp. 3812-3817, June 2004
  - [6] "TPS65120: Single-Inductor Quadruple-Output TFT LCD Power Supply", Texas Instruments, Texas Instruments Datasheet, TPS65120, 2004, <http://www.ti.com/>
  - [7] C. K. Chava and J. Silva-Martinez, "A Frequency Compensation Scheme for LDO Voltage Regulators", IEEE Transactions on circuits and systems: Regular Papers, Vol. 51, No. 6, pp. 1041-1050, 2004
  - [8] S. K. Lau, K. N. Leung and P. K. T. Mok, "Analysis of Low-Dropout Regulator Topologies for Low-Voltage Regulators", IEEE Conference on Electron Devices and Solid-State Circuits, pp. 379- 382, 2003
  - [9] G. A. Rincon-Mora and P. E. Allen, "A Low-Voltage, Low Quiescent Current, Low Drop-Out Regulator", IEEE Journal of Solid-State Circuits, Vol. 33, No. 1, pp. 36-44, 1998
  - [10] H.J. Shin, S.K. Reynolds, K.R. Wrenner, T. Rajeevakumar, S. Gowda and D.J. Pearson, "Low-Dropout On-Chip Voltage Regulator for Low-Power Circuits", IEEE Symposium on Low Power Electronics, pp. 76-77, 1994
  - [11] W. Chen, W.H. Ki and P.K.T. Mok, "Dual-Loop Feedback for Fast Low Dropout Regulators", IEEE Power Electronics Specialists Conference, Vol. 3, pp. 1265-1269, June 2001
  - [12] D. Heisley and B.Wank, "DMOS delivers dramatic performance gains to LDO regulators", EDN, Vol. 45, pp. 141-150, June 2002,
  - [13] <http://www.ednmag.com>
  - [14] T.Y . Man et al., "Developpement of single-transistor-Control LDO Based on Flipped Voltage Follower for SoC," IEEE Trans. Circuit Syst. I, vol. 55, no. 5, Jun 2008, pp. 1392-1401.
  - [15] T.Y. Man, P.K.T. Chan, "A High slew-Rate Push-Pull Output Amplifier for Low-Quiescent Current Low-DropOut regulators With Transient-Response improvement," IEEE Trans. Circuits Syst. II, vol. 54, no. 9, Sept. 2007, pp. 755-759.
  - [16] K.N. leung and P.K.T . Mok, "A Capacitor-Free CMOS Low-dropout regulator with Damping-Factor-control Frequency Compensation," IEEE J. solid-State Circuits, vol. 38, no. 10, Oct. 2003, pp. 1691-1702.
  - [17] Gianluca Giustolisi et al. "A 50-mA 1-nF Low-Voltage Low-Dropout Voltage Regulator for SoC Applications," ETRI Journal, Volume 32, Number 4, August 2010.



**Salah-ddine Krit** received the B.S. and Ph.D degrees in Microelectronics Engineering from Fes Sidi Mohammed Ben Abdellah university, Fez, Morroco. Institute in 2004 and 2009, respectively. During 2002-2008, he is also an engineer Team lead in audio and power management Integrated Circuits (ICs) Research.

Design, simulation and layout of analog and digital blocks dedicated for wireless sensor networks (WSN) and satellite communication systems using CMOS technology. He is currently a professor with Polydisciplinary Faculty of Ouarzazate, Ibn Zohr university, Agadir, Morroco. His research interests include wireless sensor Networks (Software and Hardware), microtechnology and nanotechnology for wireless communications.



**Zared kamal** received the B.S. and M.S. degrees in Electrical Engineering from faculty of science Dhar El Mehrzaz Fez in 1997 and 2004, respectively. During 2008-2010 PhD researchers in electrical engineering, 2004-2010, He is currently teacher of sciences computer.



**Hassan Qjidaa** received his M.Sc.and PhD in Applied Physics from Claude Bernard University of Lyon France in 1983 and 1987 respectively. He got the Pr. degree in Electrical Engineering from Sidi Mohammed Ben Abdellah university, Fès , Morroco 1999.

He is now an Professor in the Dept. of Physics in Sidi research interests include Very-large-scale integration (VLSI) solution, Image manuscripts Recognition, Cognitive Science, Image Processing, Computer Graphics, Pattern Recognition, Neural Networks, Human-machine Interface, Artificial Intelligence, Robotics and so on.



**Mohcine Zouak** was born in Morocco on 1963. He received the “Docteur d’Etat” degree in radar signal processing from Sidi Mohamed Ben Abdellah University, Fez (Morocco) in 1995 and Ph.D degree in electronics and informatics systems from the University of Nantes (France).

He is currently a professor with Science and Technical Faculty, Fez (Morocco), where he manages the UFR of Signals, Systems and Components. His research interests include sensors array processing, signal processing for wireless communications, and statistical signal processing. Since 2005, he has also the dean of Science and Technical Faculty, Fez.

# Automated Text Summarization Base on Lexicales Chain and graph Using of WordNet and Wikipedia Knowledge Base

Mohsen Pourvali and Mohammad Saniee Abadeh

Department of Electrical & Computer Qazvin Branch Islamic Azad University  
Qazvin, Iran

Department of Electrical and Computer Engineering at Tarbiat Modares University  
Tehran, Iran

## Abstract

The technology of automatic document summarization is maturing and may provide a solution to the information overload problem. Nowadays, document summarization plays an important role in information retrieval. With a large volume of documents, presenting the user with a summary of each document greatly facilitates the task of finding the desired documents. Document summarization is a process of automatically creating a compressed version of a given document that provides useful information to users, and multi-document summarization is to produce a summary delivering the majority of information content from a set of documents about an explicit or implicit main topic. The lexical cohesion structure of the text can be exploited to determine the importance of a sentence/phrase. Lexical chains are useful tools to analyze the lexical cohesion structure in a text. In this paper we consider the effect of the use of lexical cohesion features in Summarization, And presenting a algorithm base on the knowledge base. Our<sup>s</sup> algorithm at first find the correct sense of any word, Then constructs the lexical chains, remove Lexical chains that less score than other ,detects topics roughly from lexical chains, segments the text with respect to the topics and selects the most important sentences. The experimental results on an open benchmark datasets from DUC01 and DUC02 show that our proposed approach can improve the performance compared to sate-of-the-art summarization approaches.

**Keywords:** *text Summarization, Data Mining, Text mining, Word Sense Disambiguation*

## 1. Introduction

The technology of automatic document summarization is maturing and may provide a solution to the information overload problem. Nowadays, document summarization plays an important role in information retrieval (IR). With a large volume of documents, presenting the user with a summary of each document greatly facilitates the task of finding the desired documents. Text summarization is the process of automatically creating a compressed version of a given text that provides useful information to users, and multi-document summarization is to produce a summary

delivering the majority of information content from a set of documents about an explicit or implicit main topic [14]. Authors of the paper [10] provide the following definition for a summary: "A summary can be loosely defined as a text that is produced from one or more texts that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that. Text here is used rather loosely and can refer to speech, multimedia documents, hypertext, etc. The main goal of a summary is to present the main ideas in a document in less space. If all sentences in a text document were of equal importance, producing a summary would not be very effective, as any reduction in the size of a document would carry a proportional decrease in its informativeness. Luckily, information content in a document appears in bursts, and one can therefore distinguish between more and less informative segments. Identifying the informative segments at the expense of the rest is the main challenge in summarization". assumes a tripartite processing model distinguishing three stages: source text interpretation to obtain a source representation, source representation transformation to summary representation, and summary text generation from the summary representation. A variety of document summarization methods have been developed recently. The paper [4] reviews research on automatic summarizing over the last decade. This paper reviews salient notions and developments, and seeks to assess the state-of-the-art for this challenging natural language processing (NLP) task. The review shows that some useful summarizing for various purposes can already be done but also, not surprisingly, that there is a huge amount more to do. Sentence based extractive summarization techniques are commonly used in automatic summarization to produce extractive summaries. Systems for extractive summarization are typically based on technique for sentence extraction, and attempt to identify the set of sentences that are most important for the overall understanding of a given document. In paper [11] proposed

paragraph extraction from a document based on intra-document links between paragraphs. It yields a text relationship map (TRM) from intra-links, which indicate that the linked texts are semantically related. It proposes four strategies from the TRM: bushy path, depth-first path, segmented bushy path, augmented segmented bushy path.

In our study we focus on sentence based extractive summarization. In this way we to express that The lexical cohesion structure of the text can be exploited to determine the importance of a sentence. Eliminate the ambiguity of the word has a significant impact on the inference sentence. In this article we will show that the separation text into the inside issues by using the correct concept Noticeable effect on the summary text is created. The experimental results on an open benchmark datasets from DUC01 and DUC02 show that our proposed approach can improve the performance compared to state-of-the-art summarization approaches.

The rest of this paper is organized as follows: Section 2 introduces related works, Word sense disambiguation is presented in Section 3, clustering of the lexical chains is presented in Section 4, text segmentation base on the inner topics is presented in Section 5, The experiments and results are given in Section 6. Finally conclusion presents in section 7.

## 2. Related work

Generally speaking, the methods can be either extractive summarization or abstractive summarization. Extractive summarization involves assigning salience scores to some units (e.g. sentences, paragraphs) of the document and extracting the sentences with highest scores, while abstraction summarization (e.g. <http://www1.cs.columbia.edu/nlp/newsblaster/>) usually needs information fusion, sentence compression and reformulation [14].

Sentence extraction summarization systems take as input a collection of sentences (one or more documents) and select some subset for output into a summary. This is best treated as a sentence ranking problem, which allows for varying thresholds to meet varying summary length requirements. Most commonly, such ranking approaches use some kind of similarity or centrality metric to rank sentences for inclusion in the summary – see, for example, [1]. The centroid-based method [3] is one of the most popular extractive summarization methods. MEAD (<http://www.summarization.com/mead/>) is an implementation of the centroid-based method for either single-or-multi-document summarizing. It is based on sentence extraction. For each sentence in a cluster of related documents, MEAD computes three features and uses a linear combination of the three to determine what sentences are most salient. The three features used are

centroid score, position, and overlap with first sentence (which may happen to be the title of a document). For single-documents or (given) clusters it computes centroid topic characterizations using tf-idf-type data. It ranks candidate summary sentences by combining sentence scores against centroid, text position value, and tf-idf title/lead overlap. Sentence selection is constrained by a summary length threshold, and redundant new sentences avoided by checking cosine similarity against prior ones. In the past, extractive summarizers have been mostly based on scoring sentences in the source document. In paper [12] each document is considered as a sequence of sentences and the objective of extractive summarization is to label the sentences in the sequence with 1 and 0, where a label of 1 indicates that a sentence is a summary sentence while 0 denotes a non-summary sentence. To accomplish this task, is applied conditional random field, which is a state-of-the-art sequence labeling method. In paper [15] proposed a novel extractive approach based on manifold-ranking of sentences to query-based multi-document summarization. The proposed approach first employs the manifold-ranking process to compute the manifold-ranking score for each sentence that denotes the biased information-richness of the sentence, and then uses greedy algorithm to penalize the sentences with highest overall scores, which are deemed both informative and novel, and highly biased to the given query. The summarization techniques can be classified into two groups: supervised techniques that rely on pre-existing document-summary pairs, and unsupervised techniques, based on properties and heuristics derived from the text. Supervised extractive summarization techniques treat the summarization task as a two-class classification problem at the sentence level, where the summary sentences are positive samples while the non-summary sentences are negative samples. After representing each sentence by a vector of features, the classification function can be trained in two different manners [7]. One is in a discriminative way with well-known algorithms such as support vector machine (SVM) [16]. Many unsupervised methods have been developed for document summarization by exploiting different features and relationships of the sentences – see, for example [3] and the references therein. On the other hand, summarization task can also be categorized as either generic or query-based. A query-based summary presents the information that is most relevant to the given queries [2] and [14] while a generic summary gives an overall sense of the document's content [2], [4], [12], [14]. The QCS system (Query, Cluster, and Summarize) [2] performs the following tasks in response to a query: retrieves relevant documents; separates the retrieved documents into clusters by topic, and creates a summary for each cluster. QCS is a tool for document retrieval that presents results in a format so that a user can quickly identify a set of documents of interest. In paper [17] are developed a generic, a query-based, and a hybrid summarizer, each with



differing amounts of document context. The generic summarizer used a blend of discourse information and information obtained through traditional surface-level analysis. The query-based summarizer used only query-term information, and the hybrid summarizer used some discourse information along with query-term information. The article [18] presents a multi-document, multi-lingual, theme-based summarization system based on modeling text cohesion (story flow).

### 3. Word Sense Disambiguation

For extracting lexical chains in a document, all words and correct senses of these words should be known. Humans disambiguate words by the current context. Lexical chaining algorithms depend on an assumption, and this assumption is that correct sense of words has stronger relations with other word senses. Using this assumption, lexical chaining algorithms first try to disambiguate all word occurrences. For this reason, word sense disambiguation (WSD) is an immediate application of lexical chains and an extrinsic evaluation methodology.

#### 3.1 generating and traversing the WordNet graph

The algorithm presented in this paper is based on lexical chains therefore the system needs to deeply analyze the text. Per word has a sense based on it's position in the sentence. For instance, the word **bank** in the follow sentences has different senses: "Beautiful bank of river" and "Bank failures were a major disaster". In first sentence bank means river's coast, but in the second sentence it means economic bank. The most appropriate sense must be chosen for this word and it cause increasing the connectedness in a lexical chain. In the algorithm presented in this paper, word sense are calculated locally. In this way the best word sense is extracted. we also use WordNet as an external source for disambiguation

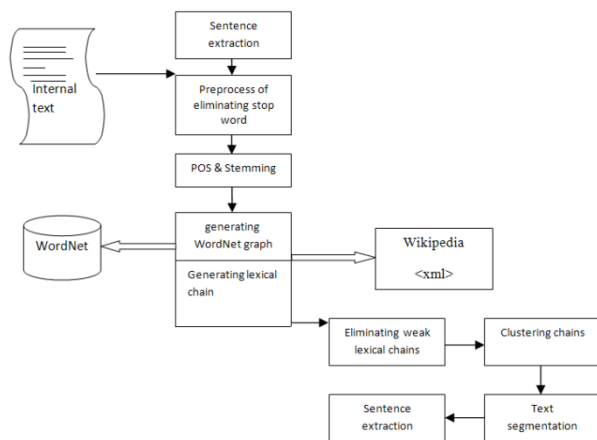


Fig. 1 Diagram of algorithm's steps

let  $w_i$  be a word in the document, and  $w_i$  have  $n$  senses  $\{w_{i_1}, w_{i_2}, \dots, w_{i_k}, \dots, w_{i_n}\}$ . In this procedure for finding the meaning of two words related locally together and placed in the same sentence, we assume all of the possible meanings and senses of per word as the first level of the traversing word tree then we process every sense in a returning algorithm. Next, we connect all the relations for that sense as it's descendants, and these descendants are generated through relations that are Hypernym, ... . We do this process in a returning manner for  $n$  levels. Next, every first level sense of the one word compare with all the first level senses of the other word. Afterwards, the numbers of equalities are considered in integer digit. The same comparison is done for another word. If there isn't any equality, for each word we choose first sense that is most common.

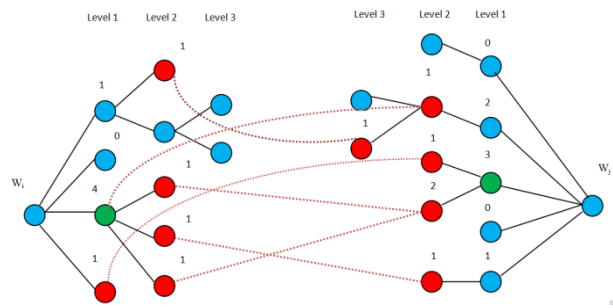


Fig. 2 Sample graph built on the 2 words

In the above figure, we illustrate the relations of the tree. The root of the tree is considered as the target word, and the first level nodes as the senses of the target words. The nodes of the second, third, ... levels are senses related with the first level nodes with Hypernym, ... relations. This tree is generated using returning functions and traversing of the tree is in the returning manner.

```

Function Hyp(ref Node t,int level)
    string[] sp
    for i = 0 to EndOfFile(wn_hyp) do
        ReadLine_From_File(wn_hyp)
        sp=Split_String_base_of(';', ',')
        if t.index == sp[1]
            tnew=Create New Nod(sp[2])
            Call Hyp(ref tnew,level-1)
            Add_New_Nod_ToList(tnew)
        end if
    end for

```

Fig. 3 Algorithm for creation WordNet graph

The above algorithm is one of the functions used for producing WordNet graph .this function is the part of the graph related with Hypernym relation .We use the great encyclopedia of Wikipedia because of the lack of special names in knowledge base of WordNet. This is done using the 3.5G XML file that is downloaded from [dumps.wikipedia.org](http://dumps.wikipedia.org) site. We have created a Xml\_Reader for this file, and then goal word abstract is extracted. Extracted abstract is used same of the Glosses of another sentence's word we use creating the graph and traversing of it just for the first ,middle ,and last sentences ,and it is useful because these sentences usually encompass concise expression of the concept of the paragraph in most of the documents .in this manner we decrease the space of interpretation and therefore the time of calculation and the space of memory because we just need to keep some highlight sentences related with each other. After clarifying the actual senses of the all words in the prominent sentences and with the similarities and relations between every pair of the words, we put them in incessant lexical chains. For example in the tree of two words, and through the traversing of the first word, we put these two words in the same lexical chain as soon as we reach the first common sense between the subordinate graph of the first word and the first level nodes of the second word .For each lexical chain  $LC_i$  ,  $w_3^1$  symbolizes that this word occur in the first sentence and the third sense of this word is chosen as the best sense. lexical chains created at first are generated from highlight sentences, and we use different algorithm for putting other words of sentences in the relevant lexical chains. in this algorithm with some changes in Lesk algorithm ,we use gloss concepts to represent similarities and differences of two words. let  $w_1$  ,  $w_2$  are two words in text .firstly we extract senses of per word in normal Lesk algorithm from knowledge base

$$s_1 \in \text{sense}(w_1) \text{ and } s_2 \in \text{sense}(w_2) \quad (1)$$

then we find overlaps between gloss concepts

$$\text{score}_{\text{lesk}}(s_1, s_2) = |\text{gloss}(s_1) \cap \text{gloss}(s_2)| \quad (2)$$

And every two concepts that have more similarities are chosen as the target words. Moreover, we use not only uni-gram (sequence of one word) overlaps , but also bi-gram (sequence of two words) overlaps .if there is one of the senses the first word in gloss concepts of the second word, we give one special score to this two senses. We do this because two concepts may have common words that are not related with their similarities and it causes increasing in scores of that two senses and makes a mistake in choosing related word as a result. Considering the word sense in gloss concept of the second word's sense, we can award an additional chance to this sense to be chosen in process of choosing words for chains from words that are not semantically related in fact.

$$\text{if}(s_1 \in \text{gloss}(s_2) \text{ or } s_2 \in \text{gloss}(s_1)) \quad (3)$$

$$\text{score}(s_1, s_2) = \text{score}_{\text{lesk}}(s_1, s_2) + \lambda$$

$\lambda$  is an additional score, and considering average existed words in sense's gloss concept and experimental tests, we find that the best value for  $\lambda$  is 5 . it is important in surveying gloss concepts to survey just existed names and existed verbs. At first, there are lexical chains generated from highlight sentences with traversing the graph, and with assuming  $LC_i$  as one of the lexical chains generated from last step and  $W_j$  as one of the other sentence's words and with using the above algorithm ,  $W_j$  is compared with members of lexical chain  $LC_i$  .if the similarity's score of  $W_j$  with one of the members of  $LC_i$  is more than threshold  $T$  ,  $W_j$  is added to  $LC_i$  and from now on, other residual words are investigated based on their similarities with members of  $LC_i$  and  $W_j$  ,too.

```

Function (Word1, Word2)
H=0 , WordInGloss = 0
For i=0 to CountOfSenseWord1
For j=0 to CountOfSenseWord2
For s=0 to CSG1[i]
For k=0 to 1
if s+k == s
    N = WSG1[s]
elseif s <> CSG1[i]
    n = WSG1[s] + “ “ + WSG2[s + k]
else break
if GlossWord2[j].Contains(n)
    H++
End if
End for
End for
if GlossWord2[j].Contains(Word1) or
    GlossWord2[i].Contains(Word1)
    WordInGloss = 5
End if
F = H + WordInGloss
ed = new edge(SenseWord1[i], SenseWord2[j], f)
AllEdge.Add(ed)
End for
End for
    
```

Fig. 4 Compare algorithm for Glosses

#### 4. Clustering Lexical Chains

After lexical chains are constructed for the text, there will be some weak lexical chains formed of single word senses.

For each lexical chain  $LC_i$ , a sentence occurrence vector  $V_i$  is formed.  $v_i = \{s_{1i}, \dots, s_{ki}, \dots, s_{ni}\}$  where  $n$  is the

number of sentences in the document. Each  $s_{k_i}$  is the number of  $LC_i$  members in the sentence  $k$ . If sentence  $k$  has 3 members of  $LC_i$  then  $s_{k_i}$  is 3. Two lexical chains  $LC_i$  and  $LC_j$  go into the same cluster if their sentence occurrence vectors  $V_i$  and  $V_j$  are similar.

Our clustering algorithm, starts from an initial cluster distribution, where each lexical chain is in its own cluster. Thus, our clustering algorithm starts with  $n$  clusters, where  $n$  is the number of lexical chains. Iteratively the most similar cluster pair is found and they are merged to form a single cluster. Clustering stops when the similarity between the most similar clusters is lower than a threshold value. for this purpose we used the well known formula from Linear Algebra:

$$Cos(\theta) = \frac{v_i + v_j}{||v_i|| ||v_j||} \quad (4)$$

In the equation  $||v_i||$  represents the Euclidean Length for the vector.

## 5. Sequence Extraction

In our algorithm, the text is segmented from the perspective of each lexical chain cluster, finding the hot spots for each topic. For each cluster, connected sequences of sentences are extracted as segments. Sentences that are cohesively connected are usually talking about the same topic. For each lexical chain cluster  $Cl_j$ , we form sequences separately. For each sentence  $S_k$ , if sentence  $S_k$  has a lexical chain member in  $Cl_j$ , a new sequence is started or the sentence is added to the sequence. If there is no cluster member in  $S_k$ , then the sequence is ended. By using this procedure, text is segmented with respect to a cluster, identifying topic concentration points. Figure 5 is an example of Text Segmentation.

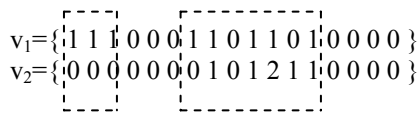


Fig. 5 example of Text Segmentation

Each sequence is scored using the formula in Equation (5).

$$Score(Sequence_i) = Score(Cl_i) * l_i * \frac{(1+SLC_i)+PLC_i}{f^2} \quad (5)$$

Where  $l_i$  is the number of sentences in the sequence,  $SLC_i$  is the number of lexical chains that starts in sequence,  $PLC_i$  is the number of lexical chains having a member in sequence, and  $f$  is the number of lexical chains in cluster. Score of the cluster score( $Cl_i$ ), is the average score of the lexical chains in the cluster. Our scoring function tries to model the connectedness of the segment using this cluster score.

## 6. Experiments and Results

In this section, we conduct experiments to test our summarization method empirically.

### 6.1 Datasets

For evaluation the performance of our methods we used two document datasets DUC01 and DUC02 and corresponding 100-word summaries generated for each of documents. The DUC01 and DUC02 are an open benchmark datasets which contain 147 and 567 documents-summary pairs from Document Understanding Conference (<http://duc.nist.gov>). We use them because they are for generic single-document extraction that we are interested in and they are well preprocessed. These datasets DUC01 and DUC02 are clustered into 30 and 59 topics, respectively. In those document datasets, stop words were removed using the stop list provided in <ftp://ftp.cs.cornell.edu/pub/smart/english.stop> and the terms were stemmed using Porter's scheme [9], which is a commonly used algorithm for word stemming in English.

### 6.2 Evaluation metrics

There are many measures that can calculate the topical similarities between two summaries. For evaluation the results we use two methods. The first one is by precision (P), recall (R) and F1-measure which are widely used in Information Retrieval. For each document, the manually extracted sentences are considered as the reference summary (denoted by  $Summ_{ref}$ ). This approach compares the candidate summary (denoted by  $Summ_{cand}$ ) with the reference summary and computes the P, R and F1-measure values as shown in formula (8) [12].

$$P = \frac{|summ_{ref} \cap summ_{cand}|}{|summ_{cand}|} \quad (6)$$

$$R = \frac{|summ_{ref} \cap summ_{cand}|}{|summ_{ref}|} \quad (7)$$

$$F_1 = \frac{2PR}{P+R} \quad (8)$$

The second measure we use the ROUGE toolkit [5], [6] for evaluation, which was adopted by DUC for automatically summarization evaluation. It has been shown that ROUGE is very effective for measuring document summarization. It measures summary quality by counting overlapping units such as the N-gram, word sequences and word pairs between the candidate summary and the reference summary. The ROUGE-N measure compares N-grams of two summaries, and counts the number of matches. The measure is defined by formula (9) [5], [6].

$$ROUGE - N = \frac{\sum_{S \in \text{sum}_{ref}} \sum_{N\text{-grams} \in S} \text{Count}_{\text{match}}(N\text{-gram})}{\sum_{S \in \text{sum}_{ref}} \sum_{N\text{-grams} \in S} \text{Count}(N\text{-gram})} \quad (9)$$

where N stands for the length of the N-gram,  $\text{Count}_{\text{match}}(N\text{-gram})$  is the maximum number of N-grams co-occurring in candidate summary and a set of reference–summaries.  $\text{Count}(N\text{-gram})$  is the number of N-grams in the reference summaries. We use two of the ROUGE metrics in the experimental results, ROUGE-1 (unigram-based) and ROUGE-2 (bigram-based).

### 6.3 Simulation strategy and parameters

The parameters of our method are set as follows: depth of tree that is created for any word,  $n=3$ ; extra value for *Lesk* algorithm,  $\lambda=5$ ; Finally, we would like to point out that algorithm was developed from scratch in C#.net 2008 platform on a Pentium Dual CPU, 1.6 GHz PC, with 512 KB cache, and 1 GB of main memory in Windows XP environment.

### 6.4 Performance evaluation and discussion

We compared our method with four methods CRF [12], NetSum [13], Manifold–Ranking [15] and SVM [16]. Tables 1 and 2 show the results of all the methods in terms ROUGE-1, ROUGE-2, and F1-measure metrics on DUC01 and DUC02 datasets, respectively. As shown in Tables 1 and 2, on DUC01 dataset, the average values of ROUGE-1, ROUGE-2 and F1 metrics of all the methods are better than on DUC02 dataset. As seen from Tables 1 and 2 Manifold–Ranking is the worst method, In the Tables 1 and 2 highlighted (bold italic) entries represent the best performing methods in terms of average evaluation metrics. Among the methods NetSum, CRF, SVM and Manifold–Ranking the best result shows NetSum.

We use relative improvement  $\frac{(\text{our method} - \text{other methods})}{\text{other methods}} \times 100$  for comparison. Compared with the best method NetSum, on DUC01 (DUC02) dataset our method improves the performance by 2.65% (3.62%), 4.26% (10.25%) and 1.81% (3.27%) in terms ROUGE-1, ROUGE-2 and F1, respectively.

Table 1:  
Average values of evaluation metrics for summarization methods (DUC01 dataset).

Methods	Av.ROUGE-1	Av.ROUGE-2	Av.F1-measure
Our method	0.47656	0.18451	0.48124
NetSum	0.46427	0.17697	0.47267
CRF	0.45512	0.17327	0.46435
SVM	0.44628	0.17018	0.45357
Manifold–Ranking	0.43359	0.16635	0.44368

Table 2:  
Average values of evaluation metrics for summarization methods (DUC02 dataset).

Methods	Av.ROUGE-1	Av.ROUGE-2	Av.F1-measure
Our method	0.46590	0.12312	0.47790
NetSum	0.44963	0.11167	0.46278
CRF	0.44006	0.10924	0.46046
SVM	0.43235	0.10867	0.43095
Manifold–Ranking	0.42325	0.10677	0.41657

## 7. Conclusion

We have attacked single document summarization. our algorithm is able to select sentences that human summarizers prefer to add to their summaries. our algorithm relies on WordNet which is theoretically domain independent, and also we have used Wikipedia for some of the words that do not exist in the WordNet. For summarization, we aimed to use more cohesion clues than other lexical chain based summarization algorithms. Our results were competitive with other summarization algorithms and achieved good results. Using co-occurrence of lexical chain members, our algorithm tries to build the bond between subject terms and the object terms in the text. With implicit segmentation, we tried to take advantage of lexical chains for text segmentation. It might be possible to use our algorithm as a text segmenter.

## References

- [1] Alguliev, R. M., & Alyguliev, R. M. (2007). Summarization of text-based documents with a determination of latent topical sections and information-rich sentences. *Automatic Control and Computer Sciences*, 41, 132–140.
- [2] Dunlavy, D. M., O’Leary, D. P., Conroy, J. M., & Schlesinger, J. D. (2007). QCS: A system for querying, clustering and summarizing documents. *Information Processing and Management*, 43, 1588–1605.
- [3] Erkan, G., & Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22, 457–479.
- [4] Jones, K. S. (2007). Automatic summarizing: The state of the art. *Information Processing and Management*, 43, 1449–1481.
- [5] Lin, C. -Y. (2004). ROUGE: A package for automatic evaluation summaries. In *Proceedings of the workshop on text summarization branches out*, (pp. 74–81). Barcelona, Spain.
- [6] Lin, C. -Y., & Hovy, E. H. (2003). Automatic evaluation of summaries using N-gram co-occurrence statistics. In *Proceedings of the 2003 conference of the north american chapter of the association for computational linguistics on human language technology (HLT-NAACL 2003)*, (pp. 71–78). Edmonton, Canada.
- [7] Mihalcea, R., & Ceylan, H. (2007). Explorations in automatic book summarization. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL 2007)*, (pp. 380–389). Prague, Czech Republic.

- [8] Navigli, R., & Lapata, M. (2010). An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation. *IEEE Computer Society*, 32.
- [9] Porter, M. (1980). An algorithm for suffix stripping. *Program*, 14, 130–137.
- [10] Radev, D., Hovy, E., & McKeown, K. (2002). Introduction to the special issue on summarization. *computational Linguistics*, 22, 399–408.
- [11] Salton, G., Singhal, A., Mitra, M., & Buckley, C. (1997). Automatic text structuring and summarization. *Information Processing and Management*, 33, 193–207.
- [12] Shen, D., Sun, J. -T., Li, H., Yang, Q., & Chen, Z. (2007). Document summarization using conditional random fields. *In Proceedings of the 20th international joint conference on artificial intelligence (IJCAI 2007)*, (pp. 2862–2867). Hyderabad, India.
- [13] Svore, K. M., Vanderwende, L., & Burges, C. J. C. Enhancing single-document summarization by combining RankNet and third-party sources. *In Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL 2007)*, (pp. 448–457). Prague, Czech Republic.
- [14] Wan, X. (2008). Using only cross-document relationships for both generic and topic-focused multi-document summarizations. *Information Retrieval*, 11, 25–49.
- [15] Wan, X., Yang, J., & Xiao, J. (2007). Manifold-ranking based topic-focused multidocument summarization. *In Proceedings of the 20th international joint conference on artificial intelligence (IJCAI 2007)*, (pp. 2903–2908). Hyderabad, India.
- [16] Yeh, J-Y., Ke, H-R., Yang, W-P., & Meng, I-H. (2005). Text summarization using a trainable summarizer and latent semantic analysis. *Information Processing and Management*, 41, 75–95.
- [17] McDonald, D. M., & Chen, H. (2006). Summary in context: Searching versus browsing. *ACM Transactions on Information Systems*, 24, 111–141.
- [18] Fung, P., & Ngai, G. (2006). One story, one flow: Hidden Markov story models for multilingual multi document summarization. *ACM Transaction on Speech and Language Processing*, 3, 1–16.
- [19] Cilibrasi, R. L., & Vitanyi, P. M. B. (2007). The Google similarity measure. *IEEE Transaction on Knowledge and Data Engineering*, 19, 370–383.

**MOHSEN POURVALI** received his B.S. degree from the Department of Computer Engineering at Razi University, in 2007. Currently, he is pursuing his M.S. degree in the Department of Electrical & Computer Qazvin University. His research areas include Data mining and Text mining.

**MOHAMMAD SANIEE ABADAH** received his B.S. degree in Computer Engineering from Isfahan University of Technology, Isfahan, Iran, in 2001, the M.S. degree in Artificial Intelligence from Iran University of Science and Technology, Tehran, Iran, in 2003 and his Ph.D. degree in Artificial Intelligence at the Department of Computer Engineering in Sharif University of Technology, Tehran, Iran in February 2008. Dr. Saniee Abadeh is currently a faculty member at the Faculty of Electrical and Computer Engineering at Tarbiat Modares University. His research has focused on developing advanced meta-heuristic algorithms for data mining and knowledge discovery purposes. His interests include data mining, bio-inspired computing, computational intelligence, evolutionary algorithms, fuzzy genetic systems and memetic algorithms.

# Designing Debugging Models for Object Oriented Systems

Sujata Khatri<sup>1</sup>, R.S.Chhillar<sup>2</sup>

<sup>1</sup> D.D.U.College University , Delhi University  
New Delhi, 110078

<sup>2</sup> D.C.S.A Maharishi Dayanand University  
Rohtak , Haryana

## Abstract

Bugs are inevitable in any software development life cycle. Most bugs are detected and removed in the testing phase. In software, we can classify bugs into two categories: (1) bugs of different severity (2) bugs of different complexity. Prior knowledge of bug distribution of different complexity in software can help project managers in allocating testing resources and tools. Various researchers have proposed models for determining the proportion of bugs present in software of different complexity but none of these models have been applied to object oriented software. Software reliability growth models have been used during later stages of testing to predict the number of latent bugs dormant in the software. Once a bug is found in the software, efforts have been made by the development team to debug it. It is found in practice that debugging may not be perfect and during removal of bugs, some new bugs may be generated and this phenomenon is called imperfect debugging. In this paper, we have developed a software reliability growth model for object oriented software system for perfect debugging in which new bugs are not generated during removal process and also by incorporating imperfect debugging, where new bugs are generated during removal process in a proportion removed bugs. Here, the proposed paper is used to assess the reliability growth of object oriented software developed under concurrent distributed development environment. We have collected bug reported data of MySQL for python. Numerical illustration has been also presented in the paper.

**Keywords:** *Open source software, software reliability growth model, and Object oriented approach.*

## 1. Introduction:

Software applications are the fastest growing trend in the virtual world and the possibilities regarding the features and functions provided by a specific application is generating tremendous interest amongst a vast number of people around the globe. As the interest grows, so does the demand for application.

Development of large software products involves several activities that need to be suitably coordinated to meet desired requirements. Meyer defines object-oriented design as "the construction of software systems as structured collections of abstract data type implementations" [1]. The emphasis on object oriented language is on defining abstraction of a model concept

related to an application domain [2]. To understand Object Oriented Programming Systems the following high level concepts must be introduced: objects, classes, inheritance, polymorphism, and dynamic binding.

Software objects are conceptually similar to physical objects; they too consist of state and related behavior. An object stores its state in fields (variables) and exposes its behavior through methods (functions). Methods operate on an object's internal state and serve as the primary mechanism for object-to-object communication [3].

The IEEE defines testing as "the process of exercising or evaluating a system or system component by manual or automated means to verify that it satisfies specified requirements or to identify differences between expected and actual results" [3]. Instead of bugs being in the software units, the complexity is now primarily in the way in which we connect the software.

The object oriented approach has been widely used for the development of closed source and open source software. In open source software, developers are also the users, meaning there by those who remove the bugs are also responsible for generating bugs [4]. Open source project has more advantage in terms of fewer bugs, better reliability, no vendor dependence, shorter development cycles, quick support and educational benefits. In the available literature, many papers address the issue of open source software [5, 6, 7, 8, and 9].

Over the last three decades various software reliability growth models have been developed but very few of them have been applied in object oriented software system. It was Kapur et al. [10] who firstly developed reliability growth model for object oriented software system. Recently, Singh et al. [19] have also developed a software reliability growth model for object oriented system which categorizes bugs into simple, hard and complex types.

During last three decades various software reliability growth models have been developed in the literature for closed source software in dealing with bug complexity. Kapur et al. [11] introduced a flexible model called the generalized Erlang SRGM by classifying the bugs in the

software system as simple, hard and complex. It is assumed that the time delay between the failure observation and its removal represent the complexity of bugs. Another model according to Kapur et al.[12], describes the implicit categorization of bugs based on the time of detection of fault. However, an SRGM should explicitly define the different types of bugs as it is expected that any type of fault can be detected at any point of testing time. Therefore, it is desired to study testing and debugging process of each type of bugs separately [13 and 14]. The mean value function of SRGM is described by the joint effect of the type of bugs present in the system. Such an approach can capture the variability in the reliability growth curve due to errors of different severity depending on the testing environment. Another model of Kapur et al. [16] describes the errors of different severity in software reliability growth model using different debugging time lag functions. Kapur et al.[17] also describe flexible software reliability growth model using a power function of testing time for defining errors of different severity. Singh et al. [18] have developed a generalized software reliability growth model which determines proportion of bugs of different complexity from open source software. Recently, Singh et al. [25] have developed a generalized software reliability growth model which determines the proportion of bugs of different complexity from open source software where software has been developed on object oriented methodology.

Due to the complexity of software systems and an incomplete understanding of software, the testing team may not be able to remove/correct the fault perfectly on observation/detection of a failure, and the original fault may remain resulting in a phenomenon known as imperfect debugging, or get replaced by another fault causing error generation. In the case of imperfect debugging, the fault content of the software remains the same; while in the case of error generation, the fault content increases as the testing progresses and removal/correction results in the introduction of new faults while removing/correcting old ones[25]. In this paper we have considered the second case where new errors are generated during removal process.

To the best of our knowledge, no research paper has addressed the issue of imperfect debugging in object oriented software systems. In this paper, we propose a Perfect and Imperfect debugging model for object oriented software system.

The rest of the paper is organized as follows: Section 2 provides model development, assumptions and formulation. Section 3 describes model validation and data analysis. And, finally, section 4 concludes the paper with a future research direction.

In this paper, we have taken actual failure data of software namely SQL for python developed under open source environment. And the development of software follows object oriented approach.

### **A. Assumptions of the Model**

Following assumptions have been taken for developing software reliability growth model for software which has been developed under open source environment using object oriented approach.

1. A finite number of test cases are prepared to ensure that the software works according to the requirements and specifications. Each test case is designed to execute a finite number of instructions.
2. The time dependent behavior of the instruction execution is represented by Exponential or Rayleigh curve.
3. The software is prone to failure due to the following causes
  - 3.1 Erroneous execution of internal variable/data of the objects. In this case we have three types of errors:
    - 3.1.1 Error due to private (local) variable/data.
    - 3.1.2 Error due to public (global) Variable/data.
    - 3.1.3 Error due to protected variable/data.
4. The failure observation/error removal phenomenon follows NHHP.
  - 5.1 No new error is introduced during removal process for perfect debugging.
  - 5.2 New errors are generated for imperfect debugging.
6. The error removal intensity per execution is proportional to the remaining errors in the software.
7. The number of executions per unit of time is proportional to the remaining number of instructions not executed.
8. The software faults are classified in to three categories.
  1. simple faults, (easy to detect and remove)
  2. hard faults (difficult to detect and remove) and
  3. complex faults. (very difficult to detect and remove)
9. We assume that accession to private, protected and public variable resulted in simple, hard and complex faults.

## **2. MODEL FORMULATION (Perfect debugging)**

The total number of instructions executed is  $E(t)$  at any given time  $t$ . These instructions cause an accession to private, protected and public variable [10 and 19]. The sum of errors (mean value function) due accession of private, protected and public variable is  $a$ .

Based on the assumptions 2 and 7, the number of instructions per unit of time can be written as

$$\frac{dE(t)}{dt} = x(t)(A - E(t)) \quad (1)$$

A is total number of instructions to be eventually executed and x(t) is the rate of instruction execution per instruction. Solving above equation we get:

$$E(t) = A \left( 1 - \exp \left( - \int_0^t x(t) dt \right) \right) \quad (2)$$

Depending upon the value of x(t), different types of instruction execution functions can be formulated. If x(t)=B i.e. instructions execution rate is independent of time then it follows exponential curve (instructions are uniformly executed) i.e.

$$E(t) = A(1 - \exp(-Bt)) \quad (3)$$

If x(t)=Bt, then it follows Rayleigh curve means instructions are no uniformly executed with respect to time.

Using assumptions (4-9), we can write the following differential equation for error removal phenomenon in case of errors due to private, protected and public variable :

$$\frac{dm_1(t)}{dt} = b(ap_1 - m_1(t)) \quad (4)$$

Where  $E_1(t)$  is the number of instructions causes an accession to private variable. Solving above differential with initial condition  $m(0)=0$ , we get  $m_1(t) = ap_1(1 - \exp(-bE_1(t)))$  (5)

Error removal equation due to accession of protected variable is

$$\frac{dm_2(t)}{dt} = \frac{b^2t}{1+bt}(ap_2 - m_2(t))$$

Where  $E_2(t)$  is the number of instructions causes an accession to protected variable. Solving above differential with initial condition  $m(0)=0$ , we get  $m_2(t) = ap_2(1 - (1 + bE_2(t))\exp(-bE_2(t)))$  (6)

Error removal equation due to accession of public variable is

$$\frac{dm_3(t)}{dt} = \frac{b^3t^2}{2 \left( 1 + bt + \frac{b^2t^2}{2} \right)} (ap_3 - m_3(t))$$

Where  $E_3(t)$  is the number of instructions causes an accession to public variable. Solving above differential with initial condition  $m(0)=0$ , we get  $m_3(t) = ap_3 \left( 1 - \left( 1 + bE_2(t) + \frac{(bE_2(t))^2}{2} \right) \exp(-bE_3(t)) \right)$  (7)

The total error removal is given as

$$m(t) = ap_1(1 - \exp(-bE_1(t))) + ap_2(1 - (1 + bE_2(t))\exp(-bE_2(t))) + ap_3 \left( 1 - \left( 1 + bE_2(t) + \frac{(bE_2(t))^2}{2} \right) \exp(-bE_3(t)) \right) \quad (8)$$

Here  $a = a(p_1 + p_2 + p_3)$  and  $E_1 = pE, E_2 = qE, \text{ and } E_3 = rE$

Here E is the total number of instructions executed due to accession of private, protected and public variables. p, q, and r is the proportion of instructions causes an accession to private, protected and public variable.  $p_1, p_2$  and  $p_3$  are proportion of faults due to accession of private, protected and public variables.

### 3. MODEL FORMULATION(Imperfect debugging)

(Case 1: Simple bugs):

$$\frac{dm_1(t)}{dt} = b_1(ap_1 + \alpha_1 m_1(t) - m_1(t))$$

Here,  $b_1$  is the bug removal rate for simple bug. a is the initial bug content in software,  $p_1$  is proportion of simple bugs,  $m_1(t)$  is the number of bugs removed due to accession of private variable and  $\alpha_1$  is bug generation rate per remaining bug

Solving above differential with initial condition  $m_1(0) = 0$  and  $m(0)=0$ , we get

$$m_1(t) = \frac{ap_1}{(1 - \alpha_1)} (1 - \exp(-(1 - \alpha_1)b_1 E_1(t))) \quad (4.1)$$



**(Case 2: Hard bugs):**

Bug removal equation due to accession of protected variable by considering bug generation during removal of bug is

$$\frac{dm_2(t)}{dt} = b_2(ap_2 + \alpha_2 m_2(t) - m_2(t))$$

Here,  $b_2$  is the bug removal rate for simple bug.  $a$  is the initial bug content in software,  $p_2$  is proportion of simple bugs,  $m_2(t)$  is the number of bugs removed due to accession of private variable and  $\alpha_2$  is bug generation rate per remaining bug.

Solving above differential with initial condition  $m_2(0) = 0$ , we get

$$m_2(t) = \frac{ap_2}{(1-\alpha_2)}(1 - \exp(-(1-\alpha_2)b_2 E_2(t))) \quad (5.1)$$

Bug removal equation due to accession of public variable by considering bug generation during removal of bug is

**(Case 3: Complex bugs):**

$$\frac{dm_3(t)}{dt} = b_3(ap_3 + \alpha_3 m_3(t) - m_3(t))$$

Here,  $b_3$  is the bug removal rate for simple bug.  $a$  is the initial bug content in software,  $p_3$  is proportion of simple bugs,  $m_3(t)$  is the number of bugs removed due to accession of private variable and  $\alpha_3$  is bug generation rate per remaining bug.

Solving above differential with initial condition  $m_3(0) = 0$ , we get

$$m_3(t) = \frac{ap_3}{(1-\alpha_3)}(1 - \exp(-(1-\alpha_3)b_3 E_3(t))) \quad (6.1)$$

The total bug removal is given as

$$m(t) = \frac{ap_1}{(1-\alpha_1)}(1 - \exp(-(1-\alpha_1)b_1 E_1(t))) + \frac{ap_2}{(1-\alpha_2)}(1 - \exp(-(1-\alpha_2)b_2 E_2(t))) + \frac{ap_3}{(1-\alpha_3)}(1 - \exp(-(1-\alpha_3)b_3 E_3(t))) \quad (7.1)$$

Here  $a = a(p_1 + p_2 + p_3)$  and  $E_1(t) = q_1 E(t)$ ,  $E_2(t) = q_2 E(t)$ , and  $E_3(t) = q_3 E(t)$  or

$$E(t) = E(t)(q_1 + q_2 + q_3)$$

Here,  $p_1$ ,  $p_2$  and  $p_3$  are proportions of bugs due to accession of private, protected and public variables.

$q_1$ ,  $q_2$  and  $q_3$  are proportions of accession due to private, protected and public variables.

## 4. MODEL VALIDATION

To verify the proposed model that determines types of fault present in the software due to accession of private, protected and public variable and proportion of instructions execution causes to accession of private, protected and public variable, we estimated the unknown parameters by using SPSS software tool.

### A. Description of Data set

**Data set-:** MySQL for Python software has been developed under open source environment [www.sourceforge.net](http://www.sourceforge.net). We collected failure data of MySQL for Python from 4/25/2001 (first bug reported) to 11/23/2009, during this period 144 bugs were reported on bug tracking system.

We have considered only valid bugs which are fixed.

We have simulated the instruction executed data with assumption that expected total number of instructions executed is 2000K and the rate of instructions execution is .003 for Rayleigh growth pattern respectively as follows in [10].

Sample of bug reported data PF MySQL for Python Software

ID	Summary	Status	Opened	Assignee	Submitter	Resolution	Priority
418713	Python 1.5.2 adds an L	Closed	4/25/2001	nobody	nobody	Wont Fix	5
419004	_mysql_timestamp_converter	Closed	4/26/2001	adustman	nobody	Fixed	5
424878	Date_or_None	Closed	5/17/2001	adustman	nobody	Fixed	5
440332	Need to #ifdef around things	Closed	7/11/2001	adustman	ads	Fixed	5
440327	setup.py configuration for my platform	Closed	7/11/2001	adustman	gimbo	Wont Fix	5
442299	core-dump. Python2.1,config_pymalloc	Closed	7/18/2001	adustman	nobody	Fixed	5
445489	Exceptions don't follow DB-API v2.0	Closed	7/28/2001	adustman	nobody	Fixed	5
464875	Limit bug in ZMySQLDA	Closed	9/25/2001	adustman	nobody	Wont Fix	5
464873	Limit bug	Closed	9/25/2001	nobody	nobody	Wont Fix	5

**A. Parameter Estimation and Comparison Results**

The performance of an SRGM is judged by its ability to fit the past software bugs and to predict satisfactorily the future behavior of the software bug removal process. Therefore, we use various comparison criteria for goodness of fit as mentioned in figure[5-8] .We have estimated the parameters of proposed model (equation 8) using SPSS tool for this data set. Parameter estimates are also shown in table [1-4].

**Parameter Estimates of My-SQL Dataset (Rayleigh) for perfect debugging.**

Table 1

Parameter Estimates of proposed model (Equation-7)	Rayleigh
a	161.905
b	.090
$p_1$	.232
$p_2$	.275
$p_3$	.493
p	.713
q	.181
r	.106

Table 2

Datasets	Comparison Results( Rayleigh)				
	$R^2$	MSE	Bias	Variation	RMSPE
MySQL for Python	.995	10.05	-0.078	3.19	3.29

**Parameter Estimates of My-SQL Dataset (Rayleigh) for Imperfect debugging**

Table 3

Parameter estimates of the proposed model equation(7)	MySQL for Python
a	174
b1	.000
b2	.473
b3	.016
p1	.000
p2	.097
p3	.903
q1	.000
q2	.774
q3	.226
$\alpha$	.041

Table 4

Datasets	Comparison Results				
	R <sup>2</sup>	MSE	Bias	Variation	RMSPE
MySQL for Python	.998	4.69304	-0.074	2.25348	2.2547

**3.5 Goodness of fit Curves**

This section describes the goodness of fit curves of different models for given data sets.

**For Perfect Debugging:**

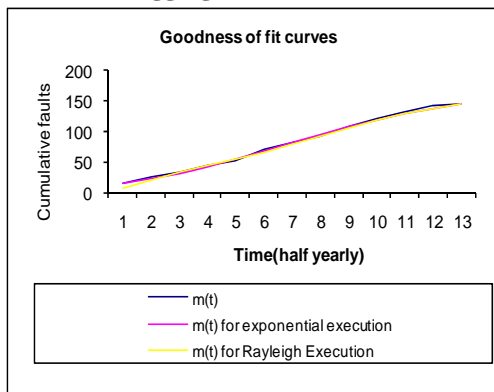


Figure 1

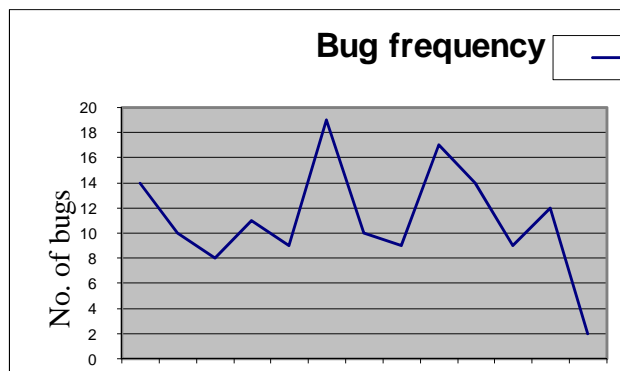


Figure 2

### For Imperfect Debugging:

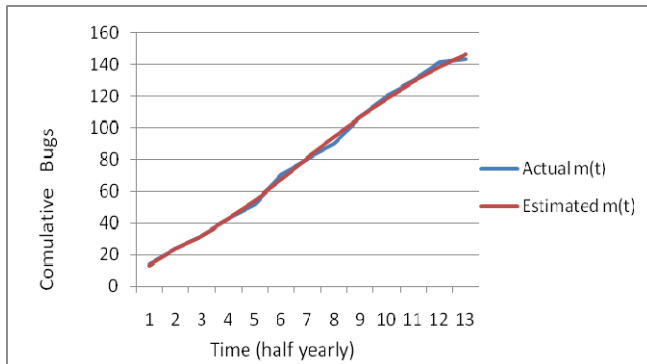


Figure 3

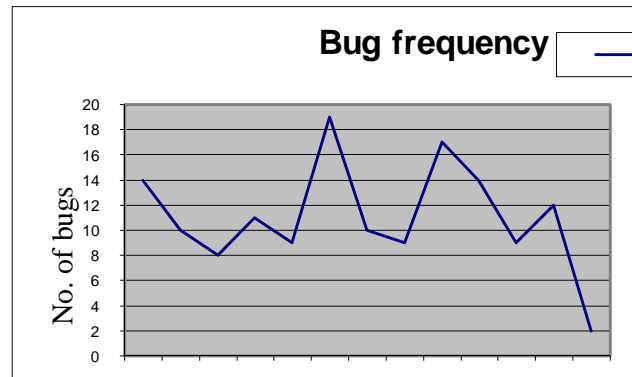


Figure 4

Figure 1, 3 gives mean value function of equation (7 and 7.1) vs time for MySql data set of Rayleigh type for Imperfect.

## 5. CONCLUSION:

Object oriented approach has become integral part of software development process. Traditional software engineering approach has been converting into object oriented software engineering. In this paper, we firstly discussed about object oriented approach. We also mentioned the main elements and advantages of using this approach. We see how accession to different variable namely private, protected and public causes an error to occur of different severity. In this paper, different modeling has been done for a failure resulting from accession to different types of variables. We have considered perfect as well as imperfect debugging occurs during bug removal. It also occurs in traditional and Object oriented development methodology. In this paper, we have proposed a software reliability growth model that determines the proportion of bug complexity in presence of perfect and imperfect debugging. The prior knowledge of distribution of bugs of different complexity will help project manager in allocation of testing efforts, tools and thus provide a better software production. We have provided the numerical results for the proposed model along with different types of growth curves depending upon their complexity.

This study can be further extended and applied on more data sets to increase confidence in the proposed model. In future, we will try to develop software reliability growth model for object oriented system by incorporating imperfect debugging and error generation by taking different models for different complexity of bugs.

## REFERENCES:

- [1] Meyer, Bertrand (1988): Object-oriented Software Construction. Prentice- Hall, New York, NY, 1988, p. 59, 62.
- [2] Binder RV: Testing object oriented software: A survey. *Journal of software testing, Verification and Reliability* 31996;6(3/4):125-252
- [3] IEEE 729-1983: Glossary of Software Engineering Terminology, *September 23, 1982*.
- [4] Gacek Cristina and Arief Budi (2004):The Many meanings of Open Source, *IEEE Software, Vol. 21, issue 1, 2004, pp.34- 40*.
- [5] Ruben van Wendel de Joode and Mark de Bruijne(2006): The organization of open source communities: Towards a Framework to Analyze the relationship between openness and reliability, *Proceedings of 39<sup>th</sup> Hawaii International Conference on System Sciences, , 2006, pp.1-6*.
- [6] Mary Paul Li, Shaw, Herbsleb Jim , Bonnie Ray, Santhanam P., Empirical Evaluation of Defect Projection Models for Widely-deployed Production Software systems, in the proceedings of the 12<sup>th</sup> *International Symposium on the production of Software Engineering (FSE-12)*, PP.263-272.
- [7] Tamura Y. and Yamada S., Optimization analysis for Reliability Assessment based on stochastic differential equation modeling for Open Source Software, *International Journal of Systems Science, Vol. 40, No.4 , 2009, pp 429- 438*.

- [8] Zhou Ying and Davis Joseph (2005): Open Source Software Reliability Model: An empirical approach, *Proceedings of the 5<sup>th</sup> WOSSE*, 2005, pp 1-6.
- [9] Singh V.B. and P.K Kapur.(2009): Measuring Reliability Growth of Open Source Software, Accepted for poster presentation in IBM-Indian Research Laboratory Collaborative Academia Research Exchange held during October 26, 2009 at IBM India Research Lab, ISID Campus, Institutional Area, Vasant Kunj, New Delhi, India.
- [10] Kapur P.K, Min Xie and Younes Said (1994): Reliability Growth Model for Object Oriented Software System, *Software Testing, Reliability and Quality Assurance*, Dec.21-22 1994., pp. 148 – 153
- [11] Kapur P.K., Younes S. and Agarwala S. (1995) ‘Generalized Erlang Software Reliability Growth Model with n types of bugs’, *ASOR Bulletin*, 14, 5-11.
- [12] Kapur P.K., Bardhan A.K., and Kumar S. (2000) :On Categorization of Errors in a Software, *Int. Journal of Kapur Management and System*, 16(1), 37-38
- [13] P.K., Bardhan A.K.; Shatnawi O.; (2002) Why Software Reliability Growth Modelling Should Define Errors of Different Severity. *Journal of the Indian Statistical Association*, Vol. 40, 2, 119-142.
- [14] Kapur P.K., Younes S and Grover P.S.; (1995), Software Reliability Growth Model with Errors of Different Severity, *Computer Science and Informatics (India)* 25(3):51-65.
- [15] Kapur P.K. Kumar Archana, Yadav Kalpana and Khatri Sunil(2007) :Software Reliability Growth Modelling for Errors of Different Severity using Change Point, *International Journal of Quality, Reliability and Safety engineering* Vol.14, No.4, pp. 311-326.
- [16] P .K Kapur. Kumar Archana Singh V.B. and Nailana F.K.(2007): On Modeling Software Reliability Growth Phenomanon for Errors of Different Severity, *In the Proceedings of National Conference on Computing for Nation Development*, Bhartiya Vidyapith’s Institute of Computer Applications and Management, New Delhi, pp.279-284, held during 23<sup>rd</sup>-24<sup>th</sup> February.
- [17] P.K., Kapur Kumar Archana, Mittal Rubina and Gupta Anu (2005): Flexible Software Reliability Growth Model Defining Errors of Different Severity, Reliability, Safety and Hazard, pp. 190-197 Narosa Publishing New Delhi.
- [18] Singh V.B., Singh O. P., Kumar.Ravi, Kapur P.K.(2010) A Generalized Software Reliability Model for Open Source Software published in proceedings of 2nd International Conference on Reliability Safety and Hazard, organized by Bhabha Atomic Research Center, Mumbai held during December, 14-16, 2010, published by IEEE Explore, pp.479-484
- [19] Singh V.B., Khatri Sujata and Kapur P.K.(2010): A Reliability Growth Model for Object Oriented Software Developed Under Concurrent Distributed Development Environment, published in proceedings of 2<sup>nd</sup> International Conference on Reliability Safety and Hazard, organized by Bhabha Atomic Research Center, Mumbai held during December, 14-16, 2010, Pp 479-484, Published by IEEE Explore. Kapur P.K., Garg R.B. and Kumar S. (1999) “Contributions to Hardware and Software Reliability”, World Scientific, Singapore.
- [20] K. Pillai and V.S.S. Nair, A Model for Software Development effort and Cost Estimation, *IEEE Transactions on Software Engineering*; vol. 23(8), 1997, pp. 485-497.
- [21] Goel, AL and Okumoto K. (1979) :Time dependent error detection rate model for software reliability and other performance Measures, *IEEE Transactions on Reliability* Vol. R-28 (3) pp.206-211.
- [22] S. Yamada, M. Ohba and S. Osaki, S-shaped Software Reliability Growth Models and their Applications, *IEEE Transactions on Reliability* R-33, 1984, PP. 169-175.
- [23] Singh V.B., Kapur P.K. and Abhishek Tandon “Measuring Reliability Growth of Software by Considering Fault Dependency, Debugging Time Lag Functions and Irregular Fluctuation” published in May issue Vol. 25, No. 3 ACM SIGSOFT Software Engineering note, 2010.
- [24] Kapur P.K., H. Pham, Anand Sameer, and Yadav Kalpana A Unified Approach for Developing Software Reliability Growth Models in the Presence of Imperfect Debugging and Error Generation, *IEEE Transaction on Reliability*, March 2011 Volume: 60 Issue: 1 On page(s): 331 - 340
- [25] Khatri Sujata, Chhilar R.S. and Singh V.B. “A Generalized Software Reliability Growth Model for Object Oriented Software” *ACM SIGSOFT*, volume 36, Issue 6, 2011.

# Designing of RF Single Balanced Mixer with a 65 nm CMOS Technology Dedicated to Low Power Consumption Wireless Applications

Raja MAHMOU<sup>1</sup>, Khalid FAITAH<sup>2</sup>

<sup>1</sup>LGECOS, ENSA Marrakech, Cadi Ayyad University, Morocco

<sup>2</sup>LGECOS, ENSA Marrakech, Cadi Ayyad University, Morocco

## Abstract

The present work consists of designing a Single Balanced Mixer (SBM) with the 65 nm CMOS technology, this for a 1.9 GHz RF channel, dedicated to wireless applications. This paper shows; the polarization chosen for this structure, models of evaluating parameters of the mixer, then simulation of the circuit in 65nm CMOS technology and comparison with previously treated.

**Keywords:** SBM Mixer, Radio Frequency, 65 nm CMOS Technology, Non-Linearity, Power Consumption.

## I. Introduction

With the multimedia's advent, the aspect of embedded systems, of RF architecture of transmission / reception channel; require a reduction in size and energy consumption with equal performances. In this stage, advanced CMOS technologies are therefore a new way now increasingly studied for the design of RF functions, the transition frequency of CMOS transistors is inversely proportional to the length of the channel, and has, as such, steady progress of lithography.

In a radio-frequency wireless, especially in a superheterodyne architecture, the frequency mixer is an indispensable module which the impact is critical on the performance of all functions [1].

circuit, calculations of evaluating parameters mixer, then a simulation of the circuit with discussion of results, and finally a potential comparison with the technologies already adopted.

## II. Modeling of evaluating mixer 'Parameters

### 2.1 Architecture of the SBM Mixer

The architecture of the mixer design, shown in Fig.2, is of the type single balanced (SBM).

This structure requires that transistors CMOS M2 and M3 must be identical. The CMOS M1 receives the RF signal and acts as the voltage / current converter. The current trail  $I_s$  is shared equally among the coupled sources M2 and M3. Thus, the  $V_{RF}$  signal varies the drain-source current of M1, and the switching operation of M2 and M3 multiplies this variation by the  $V_{LO}$  signal coming from a local oscillator. Finally, the output signal  $V_{out}$  is represented by the voltage between the drains of CMOS M2 and M3 [2].

We have chosen sizing for CMOS 65 nm technology (M1, M2, M3, M4) [3].

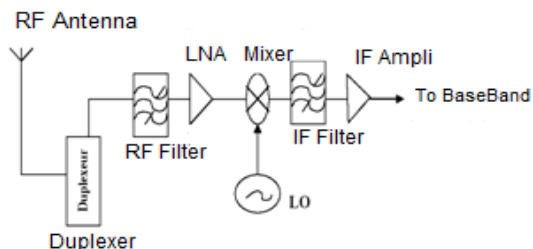


Fig.1 Functional structure of an RF receiver chain

This paper describes a work which develops the design of a single balanced mixer (SBM) with a 65 nm CMOS technology, this for a 1.9 GHz RF channel, dedicated to wireless applications, starting with the polarization of the

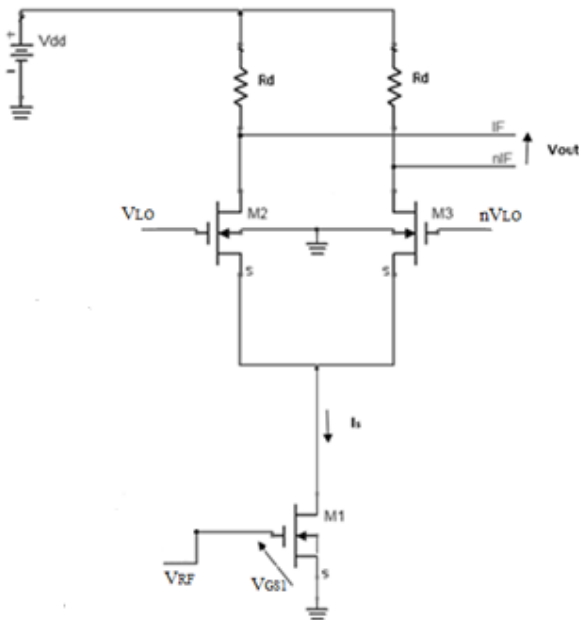


Fig.2 Polarization Diagram of a SBM

$V_{RF}$  and  $V_{LO}$  frequencies are respectively 1.9 GHz and 1.8 GHz which provides an intermediate frequency of 100 MHz. The choice of these values gives an IF frequency as agreed to meet most of the wireless networks deployed today, and operating frequency around 1 GHz, such as GSM [4].

## 2.2 Conversion Gain

The chosen architecture is an architecture single balanced with the two CMOS M2 and M3 of the differential pair are on commutation mode (Fig. 2), so the output current is controlled by the state of the VCO signal generated by the local oscillator which allows write:

$$I_{out}(t) = I_s(t) \cdot \text{signe}[V_{LO}(t)] \quad (1)$$

With  $\text{signe}[V_{LO}(t)]$  is the Fourier transform of the VCO signal:

$$\text{signe}[V_{LO}(t)] = \frac{4}{\pi} \left\{ \cos(\omega_{LO}t) - \frac{1}{3} \cos(3\omega_{LO}t) + \frac{1}{5} \cos(5\omega_{LO}t) + \dots \right\} \quad (2)$$

According to the circuit and the internal structure of CMOS M1 we have:

$$I_s(t) = g_m V_{GS1} + g_m V_{RF} \cos(\omega_{RF}t) \quad (3)$$

$g_m$  is the Transductance of the CMOS M1

$V_{GS1}$  is the bias voltage of the M1 CMOS gate, and is the DC component of the  $V_{RF}$  signal

Then we obtain:

$$I_{out}(t) = \{g_m V_{GS1} + g_m V_{RF} \cos(\omega_{RF}t)\} \frac{4}{\pi} \left\{ \cos(\omega_{LO}t) - \frac{1}{3} \cos(3\omega_{LO}t) + \frac{1}{5} \cos(5\omega_{LO}t) + \dots \right\} \quad (4)$$

And as  $V_{out}(t) = R_d I_{out}(t)$  we can write:

$$V_{out}(t) = \left\{ \frac{4g_m V_{GS1}}{\pi} R_d \cos(\omega_{LO}t) + \frac{2}{\pi} R_d g_m V_{RF} [\cos((\omega_{RF} - \omega_{LO})t) - \cos((\omega_{RF} + \omega_{LO})t)] + \dots \right\} \quad (5)$$

The conversion gain is [5]:

$$G_{conv} = \frac{V_{out_{at}(\omega_{RF}-\omega_{LO})}}{V_{RF(t)_{at}(\omega_{RF})}} = \frac{2}{\pi} R_d g_m \quad (6)$$

On the ADS2009 tool we have chosen a model for CMOS 65nm technology [7]. According to a simplified modeling of the internal structure of the transistor M1, we found  $g_m$  which is in the range of 34mA / V; this gives a theoretical conversion gain equal to 13.55dB.

## 2.3 Non-linearity

### 1 dB Compression Point

Like any electronic device with nonlinear active components, the mixer has an output power curve based on that of the entry and presenting a saturation zone. It is characterized by the 1 dB compression point, defined as the RF input power for which the conversion gain is reduced by 1 dB [6] (Fig.3).

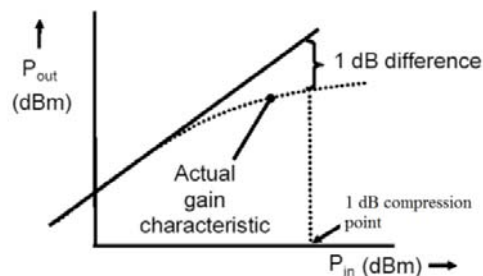


Fig.3 1dB Compression Point

### Order 3 Interception Point (IIP3)

Also called the intermodulation level of order 3, characterizes the distortion of the system, in effect: we can notice that around two useful rays  $f_1$  and  $f_2$ , can be superimposed two other very close rays ( $2 \cdot f_1 - f_2$ ) and ( $2 \cdot f_2 - f_1$ ) that can't be easily filtered out (Fig.4) [6].

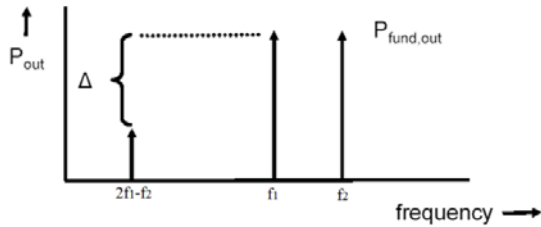


Fig.4 Order 3 interception point

IIP3 is given by the relation:

$$IIP3 = \frac{\Delta}{2} + Power_{RF} \quad (7)$$

or  $Power_{RF}$  is the input power.

▪ **Isolation**

Isolation is translated by the power coupled from one port to another. In general for a mixer, whatever its type, isolation is the most critical between RF and LO ports because of their closest frequencies and therefore difficult to filter [5].

$I_{OL,RF}$  will represent the ratio between the local oscillator's power on the channel RF and local oscillator's power injected into the mixer.

$$I_{OL,RF} = \frac{P_{OL,RF}}{P_{OL,OL}} \quad (8)$$

▪ **Noise Figure**

The noise figure NF of a mixer is defined conventionally as the degradation of signal to noise ratio between the input and the output [5]:

$$F = \frac{\frac{P_{RF}}{N_{RF}}}{\frac{P_{IF}}{N_{IF}}} \quad \text{so} \quad F = \frac{N_{IF}}{N_{RF} \cdot G_{conv}} \quad (9)$$

**III. Simulation results**

$V_{RF}$  and  $V_{LO}$  frequencies are respectively 1.9 GHz and 1.8 GHz which provides to an intermediate frequency IF of 100 MHz.

3.1 Transient signals

The Figure 5 show the shape of the IF output signal whose frequency is 100 MHz.

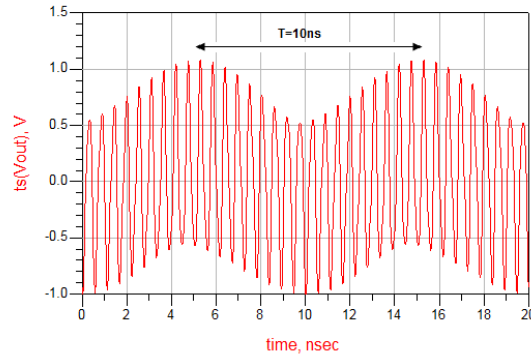


Fig.5 Time response of the output signal  $Vout(t)$

Such a chronogram represents in fact the RF carrier and the useful signal IF that we will have to restore it after an adequate filter. The following figure shows  $V(t)$ , it's the output signal  $Vout(t)$  after inserting a filter .

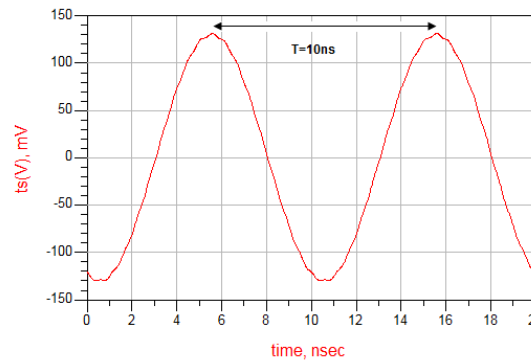


Fig.6: Response time of the output signal  $Vout(t)$  filtered

3.2 Power Consumption

DC simulation, allowed us to measure the power consumption of the mixer circuit which is 2 mW, with  $V_{dd} = 1.8V$ .

3.3 Harmonic responses

The Figures 7 and 8 show the harmonics response of order 5 of  $V_{RF}$  and  $V_{out}$  signals whose basic rays are represented respectively by 1.9 GHz and 100 MHz frequencies.



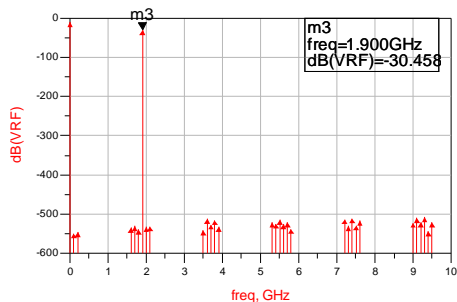


Fig.7  $V_{RF}$  harmonic response

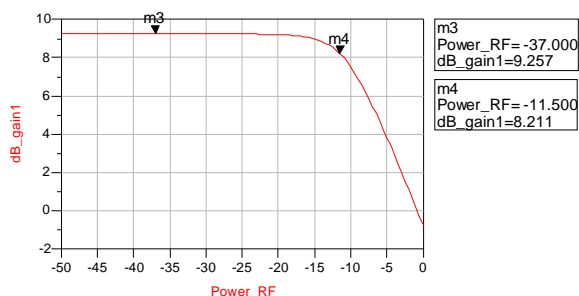


Fig.9: Gain function of the RF Power Input

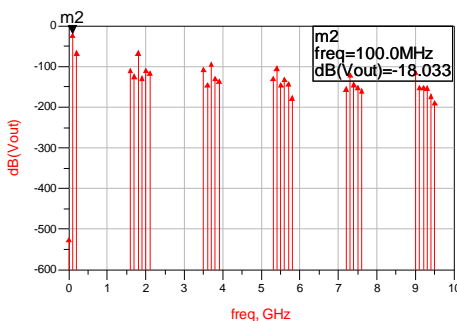


Fig.8  $V_{out}$  harmonic response

Harmonic simulation of mixer circuit results in a conversion gain (equation 6) equal to  $(-18.033\text{dB}) - (-30.458\text{dB}) = 12.425\text{dB}$  (figures 7 and 8). This reveals the importance of taking into consideration all the internal CMOS parameters which are liable to affect the results.

### 3.4 dBm Gain

We obtain the function shown in Fig.9, which represents a linear zone (horizontal) where the output is directly proportional to the input, and an area decreasing from  $-11.5\text{dBm}$  resulted by the non-linearity of the mixer circuit.

### 3.5 1 dB Compression Point

As shown in Figure 10, it's the gain value for which the output power ( $V_{out\_dBm1}$ ) does not follow its right line (Line1) with a difference of 1 dB. It corresponds well to an RF power equal to  $-11.5\text{dBm}$ .

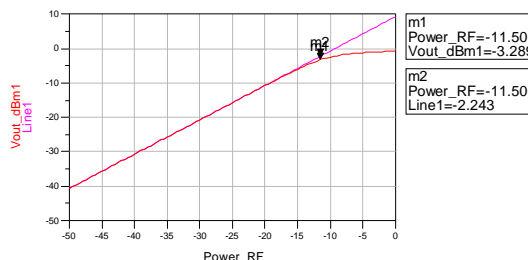


Fig.10 Response  $V_{out}$  in dBm function of the Power Input

### 3.6 Interception Point Order 3 (IIP3)

The simulation result shown in Fig.11, and (equation 7) lead to an IIP3 value in the order of  $6\text{ dBm}$ .

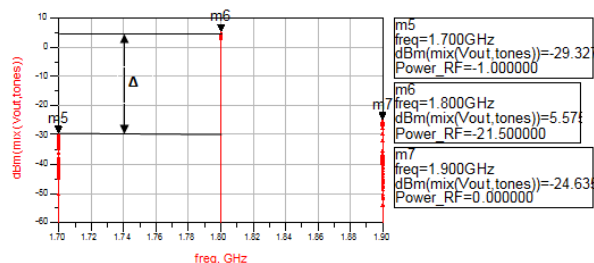


Fig.11 Interception Point IIP3

### 3.7 Measure of Isolation between ports

Isolation report (equation 8) is represented in Fig.12; it is equal to -37.704dB for 1.9 GHz RF frequency

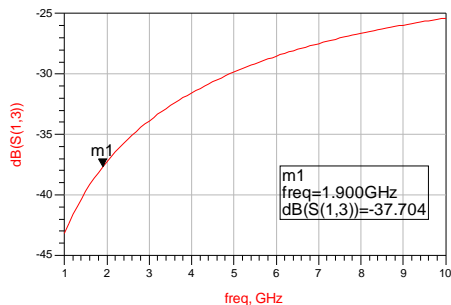


Fig.12 Isolation equal to -37.704 dB for 1.9 GHz RF frequency

### 3.8 Noise Figure

Curve noise in the input and in the output, are shown in the following figures (figures 13 and 14):

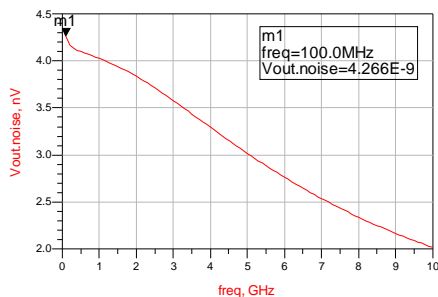


Fig.13 Noise Input

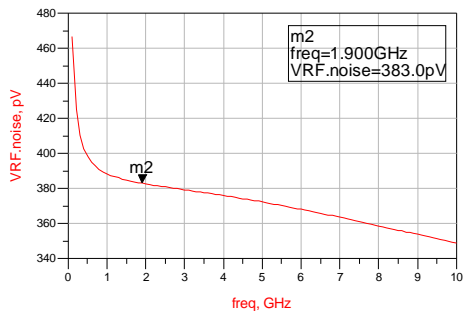


Fig.14 Noise Output

From the curves we obtain:  $NIF=4.266nV$  and  $NRF=0.383nV$  and knowing  $G_{conv} = 12.425dB$ , by the relation (9) the noise figure is 8.92dB.

## IV. Comparison of performance obtained with recent mixers

According to the simulation results found, choosing the 65nm CMOS technology with the design of other parameters (Rd, Z, .. etc..) of the SBM circuit, allowed us to achieve a very stable gain and linearity over a wide range of input power.

The performances of this mixer (SBM) are compared in the table below with that of some recent mixers:

Table 1: Comparison of Performance Obtained With Recent Mixers

Réf	Technology (μm)	RF (GHz)	CG (dB)	NF (dB)	P-1 (dBm)	IIP3 (dBm)	Pcons (mW)
[8]	0.35	0.9	1.1	-	-15.4	-3.3	7.2
[9]	0.25	2.44	-2.6	13.67	5.07	12.81	13.3
[10]	0.18	2.4	3.3	14.8	-8.98	5.46	5.6
[1]	0.18	1.9	7	8	-10	-5	3.8
<b>Proposed Circuit</b>	<b>0.065</b>	<b>1.9</b>	<b>12.42</b>	<b>8.92</b>	<b>-11.5</b>	<b>6</b>	<b>2</b>

The dynamics of an electronic circuit is defined as the power range for which the functioning is satisfactory. For lower levels, the limitation is set by the noise floor. For high levels, the limiting phenomenon is the compression. Therefore the dynamics of a mixer will be as greatest as its Intercept Point of Order 3 and its 1 dB compression point are important.

We note that the CMOS 65 nm technology SBM design with we proposed is performing well in terms of Conversion Gain, Power Consumption, levels of IIP3 point and 1 dB compression point, noise figure still acceptable.

## V. Conclusion

The research work presented in this paper is part of the overall objective; to study the feasibility of a Single Balanced Mixer (SBM) in a RF chain, dedicated to wireless applications; by 65nm CMOS technology, also to see from the simulation results, the performance of this choice compared to recent technologies, and finally to proceed to the implementation of this choice.

## References

- [1] K. Faitah, A. El Oualkadi, A. Ait Ouahman : "CMOS RF down-conversion mixer design for low-power wireless communications" ACM Ubiquity, Volume 9, Issue 24 June 17 – June 23, 2008.
- [2] M. Kramar, S. Spiegel, F. Ellinger, G. Boeck, "A Broadband Folded Gilbert-Cell CMOS Mixer", 14th IEEE International Conference on Electronics, Circuits and Systems, ICECS 2007, December 11-14, 2007. Marrakech, Morocco.
- [3] B.Martineau, "Millimeter-Wave building blocks design methodology in CMOS 65nm process using Agilent tools", ADS users's group meeting, June 16<sup>th</sup> 2009. STMicroelectronics-Crolles/Minatoc.
- [4] B. Razavi, "Design considerations for direct conversion receivers", IEEE Transactions on Circuits and Systems-II, Analog and Digital Signal Processing, vol.44, pp: 428-435 (1997).
- [5] M.Villegas, "Conception de Circuits RF et micro-ondes", Radio Communications Numeriques/2, 2<sup>ème</sup> édition, pp : 219 222, Dunod 2002-2007
- [6] K.Faitah, "Design of High-Frequency Integrated Transceiver Front-end", ENSA Marrakech, pp 29-30, 39-40, 2008-2009
- [7] T.Gneiting, "BSIM4, BSIM3v3 and BSIMSOI RF MOS Modeling", Agilent EEsof EDA Seminar, April 04, 2001 Mountain View, CA, USA.
- [8] C.F Au-Yeung and K.K.M.Cheng, "CMOS mixer Linearization by the low frequency signal injection method," IEEE MTT-S International Microwave Symposium Digest, vol 1, pp: 95-98, June (2003)
- [9] Kumar Munusamy and Zubaida Yusoff "A Highly Linear CMOS Down Conversion Double Balanced Mixer" ICSE 2006, Kuala Lumpur, Malaysia, pp: 985-990, (2006).
- [10] Hung-Che Wei, Ro-Min Weng, Chih-Lung Hsiaoand Kun-Yi Lin, " A 1.5V 2.4GHz CMOS \ Mixer With High Linearity", The 2004 IEEE Asia-Pacific Conference Circuit and Systems, pp: 6-9, Dec, (2004).

**Raja MAHMOU:** Received Professional License degree of Electrical Engineering on July 2007, then Master of Electrical Engineering on July 2009, from the Faculty of Sciences and Technics of Marrakech, at Cadi Ayyad University, Marrakech, Morocco. Now, PhD candidate on Specialty Conception of Analog Systems and RFIC at the Laboratory Of Electrical Engineering and Control Systems (LGEOS) in the National Institute of Applied Sciences (ENSA), Marrakech, Morocco, also actually a Visiting Professor of Analog Electronic, RF CMOS Designs at ENSA Marrakech.

**Dr. Khalid FAITAH:** was born in Rabat, Morocco, in June 1965. He received B.S degree in electronics from Mohamed V University of science, Rabat, Morocco in 1988 and the M.S degree in signal processing in 1997 from Hassan II University of science, Casablanca, Morocco. He received Ph.D degree in electronic at the Ibn Tofal University of science, Knitra, Morocco in 2003. In 2009 he graduated from HDR (Certificate of Accreditation to the search direction). He is now Professor at ENSA (National Institute of Applied Sciences), Department of Electrical Engineering in Cadi Ayyad University, Marrakech, Morocco where he teaches analog electronics, RF CMOS design, sensors and interface circuits and also is Responsible of the Electrical Engineering and Control Systems Laboratory, His research interests include signal integrity and analog RF CMOS design. He is the author/co-author of several publications and communications in recognized journals and international conferences.

# Analyzing the Complexity of Java Programs using Object - Oriented Software Metrics

Arti Chhikara<sup>1</sup> and R.S.Chhillar<sup>2</sup>

<sup>1</sup>Maharaja Agrasen College, Delhi, India.

<sup>2</sup>Deptt. Of Computer Sc. And Applications, Rohtak, India.

## Abstract

Object-oriented technology has rapidly become accepted as the preferred paradigm for large-scale system design. With the help of this technology we can develop software product of higher quality and lower maintenance cost. It is evident that the available traditional software metrics are inadequate for case of object-oriented software, as a result a set of new object oriented software metrics came into existence. Object Oriented Metrics are the measurement tools adapted to the Object Oriented paradigm to help manage and foster quality in software development.

Measurement of software complexity has been of great interest to researchers in software engineering for some time. Software complexity has been shown to be one of the major contributing factors to the cost of developing and maintaining software. In this research paper we investigate several object oriented metrics proposed by various researchers. These object oriented metrics are then applied to several java programs to analyze the complexity of software product.

Keywords: Object Oriented Software Development, Software Metric, Software Product, Java.

## 1. Introduction:

Object-oriented technologies reflect a natural view of the world. Object-oriented software is easier to maintain because its structure is inherently decoupled. Object Oriented Analysis and Design of software provide many benefits to both the program designer and the user. This technology promises greater programmer productivity, better quality of software and lesser maintenance cost [1,2].

OO approaches control complexity of a system by supporting hierarchical decomposition through both data and procedural abstraction [3]. However, as Brooks points out, "The complexity of software is an essential property, not an accidental one" [4]. The OO decomposition process merely helps control the inherent complexity of the problem; it does not reduce or eliminate the complexity. Measurement of the software complexity of OO systems has the potential to aid in the realization of these expected benefits. Software complexity has been shown to be

one of the major contributing factors to the cost of developing and maintaining software [6]. According to Coad and Yourdon [5], a good OO design is one that allows trade-offs of analysis, design, implementation and maintenance costs throughout the lifetime of the system so that the total lifetime costs of the system are minimized. Software complexity measurement can contribute to making these cost trade-offs in two ways. These are:

1) To provide a quantitative method for predicting how difficult it will be to design, implement, and maintain the system.

2) To provide a basis for making the cost trade-offs necessary to reduce costs over the lifetime of the system.

In this research paper different java programs are studied and object oriented software metrics are applied to them and a study of complexity is made based on the results obtained by applying object oriented metrics to different java programs.

The rest of the paper is organized as follows. Section 2 give a brief overview of object oriented metrics that we have used in our paper. Section 3 presents an example of java source code. Section 4 presents results obtained by applying object oriented metrics to java source code. Section 5 presents conclusion.

## 2. Literature Research

### 2.1 Object Oriented Metrics

One of the most widely referenced sets of object-oriented software metrics has been proposed by Chidamber and Kemerer [7,11]. At the 1991 Object Oriented Programming Systems, Languages and Applications conference (OOPSLA), Shyam Chidamber and Chris Kemerer presented a paper [7] outlining six metrics for use with object-oriented programming languages. The metrics used in this study are given below:

1 **Weighted Method per Class(WMC)**: WMC is defined as the sum of the complexities of all methods of a class. If there are n methods of complexities

$c_1, c_2, \dots, c_n$  are defined for a class C. The specific complexity metric that is chosen should be normalized so that nominal complexity for a method takes on a value of 1.0 [16].

$$WMC = \sum c_i \quad \text{for } i=1 \text{ to } n$$

The number of methods and their complexity are reasonable indicators of the amount of effort required to implement and test a class. In addition, the larger the number of methods, the more complex is the inheritance tree (all subclasses inherit the methods of their parents). Finally, as the number of methods grows for a given class, it is likely to become more and more application specific, thereby limiting potential reuse. For all of these reasons, WMC should be kept as low as is reasonable [16].

**2 Depth of Inheritance Tree(DIT):** This metric is “a measure of how many ancestor classes can potentially affect this class.” [7,10] The deeper a class is in the inheritance the more behavior it is likely to inherit from its superclasses. Deep inheritance trees are indicative of complex designs. This metric is useful as a design aid in designing classes that make use of inherited methods.

**3 Number of Children(NOC):** The NOC is the number of immediate subclasses in the hierarchy. High NOC indicates high reuse. But, if there are a large number of children of a class, then the abstraction level of that parent class is reduced. If a class has too many children, it may indicate misuse of sub-classing. The number of children gives an idea of the potential influence a class has on the design. If a class has a large number of children, it may require more testing[7,10].

**4 Response For a Class (RFC):** This metric is a count of all member functions called by any member function in the class being measured. Member functions in the class and member functions of other classes are both counted equally. It is “considered a measure of attributes of an object. Since it specifically includes methods called from outside the object, it is also a measure of communication between objects.” [7]

Several studies have been conducted to validate CK’s metrics. Their metrics have been criticized, specially the LCOM metric, for being too ambiguous for practical applications and for not being language independent [12]. Basili et al. [13] presented the results of an empirical validation of CK’s metrics.

Tang et al. [14] validated CK’s metric suit using real time systems.

Li, et al. have also empirically evaluated C&K’s metrics as being predictors of maintenance effort [15]. In addition, Li, et al. [15] proposed new metrics that were used in their study including:

**5 Message passing coupling:** The Message Passing Coupling metric measures the number of method calls defined in methods of a class to methods in other classes, and therefore the dependency of local methods to methods implemented by other classes. It allows for conclusions on the message passing (method calls) between objects of the involved classes. This allows for conclusions on reusability, maintenance and testing effort.

**6 Data abstraction coupling:** Data abstraction coupling is a count of total number of instances of other classes within a given class. It is the count of total number of external classes the given classes uses.

**7 Number of local subunits:** The number of local subunits is the total number of functions and procedures defined for a class. Classes with large number of operations are harder to maintain and are more fault prone.

Morris [8,9] in 1989 made some important observations on OO code and proposed candidate metrics for productivity measurement:

**8 Inheritance Dependencies:** This metric is intended to reflect characteristics of the inheritance tree. Morris suggests that “it may be possible to determine a range of values within which the inheritance tree depth should be maintained.”[9]:

This metric is calculated using the following equation:

$$\text{Inheritance tree depth} = \max(\text{inheritance tree path length})$$

**9 Factoring Effectiveness:** Morris states that “inheritance hierarchies are optimized via a process called factoring. The purpose of factoring is to minimize the number of locations within an inheritance hierarchy in which a particular method is implemented.”[9]

It is calculated as below:

$$\text{Factoring Effectiveness} = \text{Number of unique methods} / \text{Total number of methods}$$

**10 Reuse Ratio:** The reuse ratio, RR is given by [18]:

Reuse Ratio = Number of Superclasses/Total number of Classes

**11 Specialization Index:** Specialization Ratio (SR): Specialization ratio, SR is given as [18]:

Specialization Index = Number of Subclasses/Number of Superclasses

### 3. Definition of Metric

To better define and understand how these metrics are calculated using java source code example is used.

#### 3.1: Java source code[1,2]

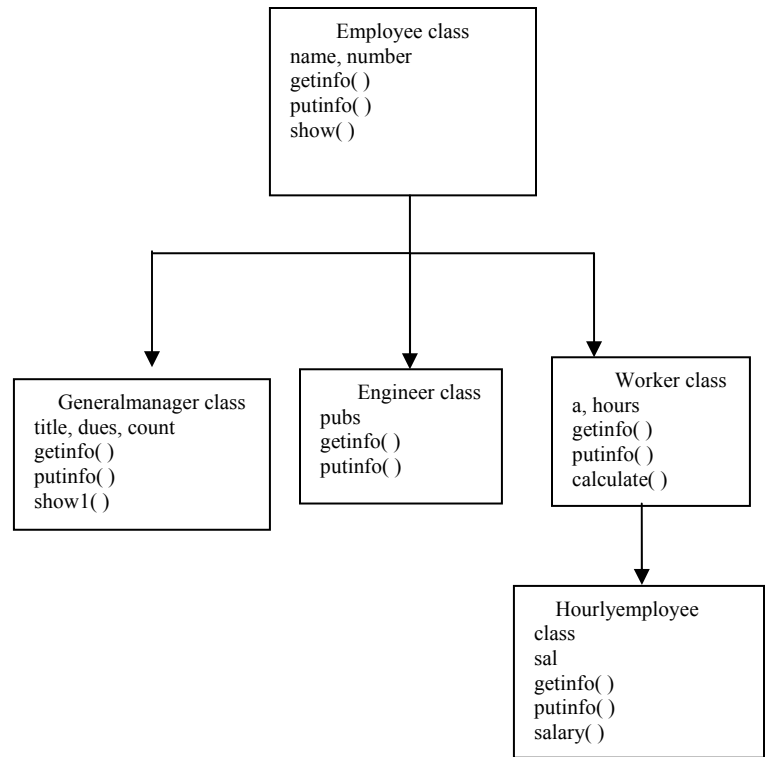
```
import java.io.*;
public class employee
//employee class
{
    DataInputStream in=new
DataInputStream(System.in);
    private String name;
//employee name
    public int number;
//employee number
    public void getinfo()
    {
        System.out.println("Enter name:");
        name= in.readLine();
        System.out.println ("enter number :");
        number=Integer.parseInt(in.readLine());
    }
    public void putinfo()
    {
        System.out.println("The name is:" +name);
        System.out.println ("Number=" +number);
    }
    public void show()
    {
        System.out.println("End of Employee Class");
    }
}
public class generalmanager extends employee
//generalmanager class
{
    private String title ;
    private double dues ;
    private int count;
    count = total
    public void getinfo()
    {
        super.getinfo();
        System.out.println("enter title :");
```

```
        title=in.readLine();
        Console.WriteLine("enter golf club dues:");
        dues=double.parseDouble(in.readLine());
    }
    public void putinfo()
    {
        super.putinfo();
        System.out.println(count);
        System.out.println ("title:" +title);
        System.out.println("dues:" +dues);
    }
    public void show1()
    {
        System.out.println("End of manager class");
    }
}
public class engineer extends employee
// engineer class
{
    private int pubs ;
    public void getinfo()
    {
        super.getdata();
        System.out.println ("enter number of pubs:");
        pubs=Integer.parseInt(in.readLine());
    }
    public void putinfo()
    {
        super.putinfo();
        System.out.println ("number of pubs:" +pubs);
    }
}
public class worker extends employee
// worker class
{
    private int a;
    public int hours;
    public void getinfo()
    {
        super.getdata();
        System.out.println ("Enter number of hours:");
        hours=Integer.parseInt(in.readLine());
    }
    public void calculate( )
    {
        int total=0;
        total = LEN*40;
    }
    public void putinfo()
    {
        super.putinfo();
        System.out.println ("number of hours :" +hours) ;
        System.out.println ("Total:" +total);
    }
}
public class hourlyemployee extends worker
//hourlyemployee class
```

```

{
private double sal;
public void getinfo()
{
super.getinfo();
System.out.println ("enter number of hours.");
hours=Integer.parseInt(in.readLine());
}
public void salary()
{
sal=super.hours*250;
// calling superclass instance variable
}
public void putinfo()
{
super.putinfo();
System.out.println ("The salary is: "+sal);
}
public static void main(String args[])
//main method
{
generalmanager m1 = new generalmanager();
generalmanager m2 = new generalmanager();
technician s1= new scientist();
worker L1 = new worker();
hourlyemployee h1 = new hourlyemployee();
System.out.println ("Enter data for manager 1");
//get data for several employees
m1.getinfo();
System.out.println ("Enter data for manager 2");
m2.getinfo();
System.out.println ("Enter data for scientist 1");
s1.getinfo();
System.out.println ("Enter data for laborer 1");
L1.getinfo();
System.out.println ("Enter data for hourlyemployee
1");
h1.getinfo();
System.out.println ("Data on manager 1");
m1.putinfo();
System.out.println ("Data on manager 2 ");
m2.putinfo();
System.out.println ("Data on scientist 1");
s1.putinfo();
System.out.println ("Data on Laborer 1");
L1.putinfo();
System.out.println ("Data on hourly employee");
h1.putinfo();
}
    
```

**3.2 Class Diagram for java source code:**



**Fig 1: Class Diagram**

**3.3 Object Oriented Software Metrics Applied on Example 1:**

**1. WMC (Weighted Method per Class):** WMC is calculated by counting the number of methods in each class [4].

Metric	Employee class	Manager Class	Engineer class	Laborer class	Hourlyemployee class
WMC	3	3	2	3	3

**2. RFC (Response for a Class):** The RFC is the number of functions or procedures that can be potentially be executed in a class. Specifically, this is the number of operations directly invoked by member operations in a class plus the number operations themselves [4].

Metri c	Emplo ee class	Manag er Class	Engin er class	Work er class	Hourlyemplo yee class
RFC	3	5	4	5	7

**3. DIT (Depth of Inheritance tree):** The depth of inheritance is defined to be the level of the class in the inheritance hierarchy, with the root class being Zero [4].

Metri c	Employ ee class	Manag er Class	Engine er class	Labor er class	Hourlyemplo yee class
DIT	0	1	1	1	2

**4. NOC (Number of Children):** The number of children is the number of direct descendents for a class [4].

Metri c	Employ ee class	Manag er Class	Engine er class	Labor er class	Hourlyemplo yee class
NOC	3	0	0	1	0

**5. MPC (Message Passing Coupling):** Message Passing coupling is the count of total number of function and procedure calls made to external units [7].

Metri c	Employ ee class	Manag er Class	Engine er class	Labor er class	Hourlyemplo yee class
MPC	0	2	2	2	4

**6. DAC (Data Abstraction Coupling):** Data Abstraction coupling is the count of total number of instances of other classes within a given class [7].

Metri c	Employ ee class	Manag er Class	Engine er class	Labor er class	Hourlyemplo yee class
DAC	0	1	0	1	0

**7. NUS (Number of Subunits):** The number of subunit is the total number of functions and procedures defined for the class [7].

Metri c	Employ ee class	Manag er Class	Engine er class	Labor er class	Hourlyemplo yee class
NUS	3	3	2	3	3

**8. Inheritance dependencies(ID):** This metric is calculated using the following equation:  
 Inheritance tree depth=max(inheritance tree path length)  
 So referring the above class diagram  
 Inheritance tree depth = 3

**9. Factoring effectiveness(FE):** This metric is calculated using the following equation:  
 Factoring effectiveness = No. of unique methods / Total no. of methods  
 = 4/14  
 = 0.29

**10. Specialization index(SI):** This metric is calculated using the following equation:  
 Specialization index= Total no. of subclasses / total no. of superclasses  
 From the above class diagram  
 Specialization index =5/2  
 = 2.5

**11. Reuse ratio(RR):** This metric is calculated using the following equation:  
 Reuse ratio=Total no. of superclasses / Total no of classes  
 Referring above class diagram  
 Reuse ratio =2/5  
 =0.4

#### 4. Study of the complexity of Java programs

These metrics were calculated and tested on 20 Java programs and following results are obtained.

Table 1 shows the metric value for 20 java programs and Table 2 shows the statistical values calculated for the metric values obtained from 20 java programs



**Table 1: Metric Values Calculated for JAVA Programs**

Metrics Type	Program Number																			
	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20
WMC	3.00	2.25	1.65	2.00	2.00	1.25	2.25	2.00	2.00	1.65	3.33	1.50	2.00	2.00	1.00	1.67	2.00	3.33	2.00	2.00
RFC	2.00	3.00	3.33	2.00	3.33	3.33	2.67	3.00	4.48	1.50	3.00	3.33	2.50	2.00	3.33	3.33	2.00	4.10	3.00	2.00
DIT	2.00	1.00	1.00	1.00	0.50	2.00	0.50	0.33	1.00	1.00	0.75	0.50	0.50	1.00	1.00	2.25	0.33	0.33	0.50	0.50
NOC	2.00	0.75	0.50	1.00	0.50	1.50	0.50	0.50	0.50	0.65	0.65	1.00	1.00	1.00	0.75	1.75	0.67	0.50	0.50	0.50
MPC	2.00	0.33	0.20	0.33	0.00	0.00	0.20	0.00	0.00	0.00	0.50	0.50	0.33	0.33	0.33	0.00	0.00	0.33	0.00	0.00
DAC	0.30	0.00	0.00	0.40	0.67	0.00	0.00	0.00	0.33	0.50	0.67	0.50	0.50	0.67	0.00	0.33	0.33	0.30	0.30	0.40
NUS	3.00	2.00	1.65	1.65	2.00	2.00	2.00	1.67	1.67	1.33	1.50	1.50	1.50	2.50	2.00	1.67	1.67	2.50	2.00	2.00
ID	2.00	1.00	0.50	0.50	1.00	0.33	2.00	1.00	1.00	1.00	0.50	0.50	1.00	1.00	2.00	2.25	2.00	0.50	0.50	0.33
FE	0.50	0.50	0.30	0.30	0.50	0.67	0.67	0.67	1.25	0.50	0.33	0.67	0.67	0.33	0.33	0.33	0.50	0.67	0.50	0.50
SI	2.00	2.00	1.00	1.00	1.00	3.00	2.00	2.00	1.00	1.00	2.00	3.00	1.00	1.00	1.00		2.00	3.00	2.00	2.00
RR	0.25	0.33	0.25	0.25	0.50	0.50	0.33	0.33	0.30	0.33	0.25	0.75	0.50	0.50	0.30	0.25	0.25	0.30	0.50	0.30

**Table2: Statistical Values Calculated for JAVA Programs**

Metric Type	Minimum	Maximum	Mean	Median	Stand. Deviation
WMC	1.00	3.33	2.04	2.00	0.59
RFC	1.50	4.48	2.86	3.00	0.77
DIT	0.33	2.25	0.89	0.87	0.57
NOC	0.50	2.00	0.83	0.66	0.44
MPC	0.00	2.00	0.26	0.20	0.44
DAC	0.00	0.67	0.31	0.33	0.23
NUS	1.33	3.00	1.89	1.83	0.40
ID	0.33	2.25	1.04	1.00	0.64
FE	0.30	1.25	0.53	0.50	0.21
SI	1.00	3.00	1.75	2.00	0.71
RR	0.25	0.75	0.36	0.31	0.13

**After analyzing the Table 1 and Table 2 following points are observed regarding complexity of the java programs**

**1** Weighted Method per Class metric predicts time and effort that is required to build and maintain a class. A high value of WMC has been found to lead to more faults. Classes with large number of methods are likely to be more application specific, limiting the possibility of reuse. A study of 20 java programs suggest that an increase in the average WMC increases the complexity and decreases quality. As programs with large number of methods are more prone to bugs and complex to understand.

**2** The RFC metric is the count of the set of all methods that can be invoked in response to a message to an object of the class or by some methods in the class. This includes all methods accessible within the class hierarchy. This metric looks at the combination of the complexity of a class through the number of methods and the amount of communication with other classes. The larger the number of methods that can be invoked from a class through messages, the greater the complexity of the class. From our study we found that Java programs are less complex as the mean value of this metric is low for java programs.

**3** The depth of a class within the inheritance hierarchy is the maximum number of steps from the class node to the root of the tree and is measured by the number of ancestor classes. The deeper a class is in the hierarchy, the more methods it is likely to inherit, making it more complex. Deeper trees constitute greater design complexity, since more methods and classes are involved, but at the same time reusability also get increase due to inheritance. Java programs have intermediate value for DIT metric.

**4** The number of children is the number of immediate subclasses subordinate to a class in the hierarchy. It is an indicator of the potential influence a class can have on the design and on the system. The greater the number of children, the greater the likelihood of improper abstraction of the parent and may be a case of misuse of subclassing. However high NOC indicates high reuse, since inheritance is a form of reuse.. A class with many children may also require more testing. High NOC has been found to indicate fewer faults. This may be due to high reuse, which is desired. In Java the value of this metric depends on program to program. All classes do not have the same number of sub-classes. However, it is observed that for better results, classes higher up in the hierarchy should have more sub-classes than those lower down.

**5** Message passing coupling metric measures the numbers of messages passing among objects of the class. A larger number indicates increased coupling between this class and other classes in the system. This makes the classes more dependent on each other which increases the overall complexity of the system and makes the class more difficult to change. The assumption behind this metric is that classes interacting with many other classes are harder to understand and maintain. When we applied object oriented metrics on several java programs, we observed that the value of Message Passing Coupling (MPC) metric is low for java programs.

**6** Data Abstraction Coupling metric measures the coupling complexity caused by Abstract Data Types (ADTs). This metric is concerned with the coupling between classes representing a major aspect of the object oriented design, since the reuse degree, the maintenance and testing effort for a class are decisively influenced by the coupling level between classes. It is the count of total number of external classes the given classes uses. Software complexity increases with increasing DAC.As Java is an object oriented language so data is given more importance than procedures. Data is hidden from the outside world. The value of this metric is low for java programs.

**7** The Number of Subunit metric is the total number of functions and procedures defined for the class. As the number of functions and procedures grow, class become more fault prone. The complexity also get increase with increase value of local subunits metric. The value of this metric is found to be low for java programs.

**8** Inheritance Dependencies metric is intended to reflect characteristics of the inheritance tree. Morris suggests that “it may be possible to determine a range of values within which the inheritance tree depth should be maintained. inheritance tree depth is likely to be more favorable than breadth in terms of reusability via inheritance. However, A deeper tree is more difficult to test than a broader one. The greater the value of this metric, more will be the complexity of programs. Comprehensibility may be diminished with a large number of inheritance layers.

**9** Morris states that “inheritance hierarchies are optimized via a process called factoring. The purpose of factoring is to minimize the number of locations within an inheritance hierarchy in which a particular method is implemented.”[9] Highly factored applications are more reliable for reasons similar to

those that argue that such applications are more maintainable. The smaller the number of implementation locations for the average task, the less likely that errors were made during coding. The more highly factored an inheritance hierarchy is the greatest degree to which method reuse occurs. The more highly factored an application is, the smaller the number of implementation locations for the average method.

**10** Specialization Index metric measures the extent to which subclasses override their ancestors classes. This index is the ratio between the number of overridden methods and total number of methods in a Class, weighted by the depth of inheritance for this class. This metric was developed specifically to capture the point that classes are structured in hierarchy which reuse code and specialize code of their superclasses. It is well-defined, not ambiguous and easy to calculate. However, it is missing theoretical and empirical validation. It is commonly accepted that the more the Specialization Index is elevated, the more difficult is the class to maintain. The value of this metric is high for java programs, as java classes are more usable.

**11** Reuse ratio measures reuse via inheritance. A high value of this metric indicates a deep class hierarchy with high reuse. Reuse ratio is the percentage of classes that are derived from. Reuse ratio varies in the range  $\{0,1\}$ . When the value of this metric is zero, there is no inheritance. As the value of this metric approaches 1, the inheritance tree deepens in a chain form with exactly one root and one leaf. When this metric is applied to several java programs we got intermediate results.

## 5 Conclusion and Future Work

The primary objective of this study was to investigate the applicability of Object-Oriented software metrics to measure the complexity of a Java software application. Complexity of Java applications can be evaluated at several dimensions (Size, method, class, inheritance, cohesion etc) using a variety of available software metrics from Software Engineering Domain. In this research paper we have presented a set of eleven well established object-oriented metrics that can be used to rank programs on their complexity values, to assess testability and maintainability of the programs. From this study we conclude that there should be a compromise among internal software attributes in order to maintain a high degree of reusability while keeping the degree of complexity and coupling as low as possible.

However it is still insufficient, needs further in depth study and future work will focus on empirical validation of object oriented metrics in multi languages environment. But we still expect that our analysis can be used as a reference by software developers for building a fault free, reliable, and easy to maintain software product in Java

## References:

- [1] Patrick Naughton & Herbert Schildt "java: The complete reference", McGraw-Hill Professional, UK, 2008.
- [2] Er. V.K. Jain. "The Complete Guide to java programming", First Edition, 2001.
- [3] G. Booch, Object-Oriented Design with Applications (The Benjamin/Cummings Publishing Company, Redwood City, CA, 1991; ISBN: 0-8053-0091-0).
- [4] F.P. Brooks, No Silver Bullets: Essence and Accidents of Software Engineering, Computer, Vol. 20, No. 4 (Apr 1987) 10-19.
- [5] P. Coad and E. Yourdon, Object-Oriented Design (Yourdon Press, Englewood Cliffs, NJ, 1991; ISBN: 0-13-630070-7).
- [6] R.B. Grady, Practical Software Metrics for Project Management and Process Improvement (Prentice Hall, Englewood Cliffs, NJ, 1992; ISBN: 0-13-720384-5).
- [7] S. Chidamber, and C. Kemerer, "Towards a Metrics Suite for Object Oriented Design," Object Oriented Programming Systems, Languages and Applications (OOPSLA), Vol 10, 1991, pp 197-211
- [8] Michael W. Cohn, William S. Junk, "Empirical Evaluation of a Proposed Set of Metrics for Determining Class Complexity in Object-Oriented Code", A Thesis, College of Graduate Studies University of Idaho, April 1994
- [9] K. Morris, "Metrics for Object-oriented Software Development Environments," Masters Thesis, MIT, 1989.
- [10] Chidamber, S. and Kemerer, C." A Metrics Suite for Object Oriented Design", IEEE Transactions on Software Engineering, vol. 20, no. 6, pp. 476-493, 1994.

[11]Chidamber, S., Darcy, D., Kemerer, C.” Managerial use of Metrics for Object Oriented Software”: an Exploratory Analysis, IEEE Transaction on Software Engineering, vol. 24, no. 8, pp. 629-639,1998.

[12] Churcher, N.I. and M.J. Shepperd, “Towards a Conceptual Framework for Object-Oriented Metrics,” ACM Software Engineering Notes, vol. 20, no. 2, April 1995,pp. 69–76.

[13] Basli VR, Briand LC, Melo WL. “A validation of object oriented design metrics as quality indicators”. Technical Report, University of Maryland, Department of Computer Science,1-24, 1995.

[14] Tang MH, Kao MH. “An empirical study on object-oriented metrics”. Proceedings 23<sup>rd</sup> Annual

International Computer Software and Application Conference. IEEE Computer Society, 242-249,1999.

[15] Li. W. “Another Metric suit for object-oriented programming”. The journal of system and software 44(2),155-162,1998.

[16] Roger S. Pressman: Software Engineering, A practioner’s Approach, Fifth Edition,2001.

[17] R. Kolewe, “Metrics in Object-Oriented Design and Programming,” Software Development, Vol. 1, No. 4, October 1993, pp. 53-62.

[18] Jacobson Ivar : Object Oriented Software Engineering: A Use Case Driven Approach, Addison-Wesley Publishing Company,1993

# Semantic Malware Detection by Deploying Graph Mining

Fatemeh Karbalaie<sup>1</sup>, Ashkan Sami<sup>2</sup> and Mansour Ahmadi<sup>3</sup>

<sup>1</sup>CSE&IT Department, Shiraz University  
Shiraz, Iran

<sup>2</sup>CSE&IT Department, Shiraz University  
Shiraz, Iran

<sup>3</sup>Young Researchers Club, Shiraz Branch, Islamic Azad University  
Shiraz, Iran

## Abstract

Today malware is a serious threat to our society. Several researchers are studying detection and mitigation of malware threats. On the other hand malware authors try to use obfuscation techniques for evading detection. Unfortunately usual approach (e.g., antivirus software) use signature based method which can easily be evaded. For addressing these shortcomings dynamic methods have been introduced. The aim of dynamic methods is to detect the semantic of malware family. Obfuscation of semantic based method is too difficult and results of these methods are promising. However deploying semantic based methods for real time detection have several complications. Current semantic methods are too time-consuming and usually need a robust virtual machine to obtain the behavior. In this paper we present an automatic detection method based on graph mining techniques with near optimal detection rate. That is 96.6% accuracy and only 3.4% false positive. In our method, first the malware is analyzed in a virtual machine environment to observe its semantic. A graph representation of malware behavior is constructed. The representation is based on relationships between system calls and allows rearrangement of system calls. Graph is used for representing the behavior of application because graph, especially labeled graph, can be used to model lots of complicated relation between data. At the next step we mine information graph and extract the most discriminative graphs that separate malware from benign. Finally, a classification method is used and the mentioned accuracy was obtained.

**Keywords:** *Semantic, Malware Detection, System call, frequent sub graph, labeled graph, subgraph isomorphism.*

## 1. Introduction

"Malware" is an abbreviation for 'malicious software' and is typically used as a catch-all term to refer to any software or program that damages computer systems or destroys valuable information stored in computers. Typical examples include viruses, worms, trojans, and spyware. Malware may be propagated using spam, may also be used to send spam, may take advantage of bugs, and may be used to mount DoS attacks. Recently the threat of malware has acquired an economic dimension as attackers benefit financially from compromised machines (e.g., by selling hosts as email relays to spammers) [1]. These considerations illustrate that addressing the problem of malware is necessary for improving computer security. Computer security is necessary to our society's critical infrastructure. Historically, detection tools such as signature based detection methods have performed poorly, particularly when facing previously unknown malware programs, novel variants of existing ones and polymorphic/metamorphic malware. An important problem is that many of detection techniques rely on ineffective models. Ineffective models are models that do not capture natural properties of a malicious program and its actions but merely pick up artifacts of a specific malware instance. As a result, they can be easily evaded. Most of these models capture the sequence of system calls that a

specific malware program executes. The defect of these methods is that, when these system calls are independent, it is easy to change their order or add irrelevant calls, thus evading the captured sequence.

Today for above mentioned problems, researchers propose ways to capture the malicious behavior that characterizes a malware program. On one hand, some detectors [2, 3, 4] use sophisticated static analysis to identify the code that is semantically equivalent to a malware template. These actual semantic of program is unaffected by obfuscation, but at the other hand static analysis suffer from some limitation such as difficulty of static binary analysis, high cost of doing such analysis and the low speed in scanning large number of files [5]. In this paper we propose a novel and near optimal malware detection approach base on dynamic analysis. Also dynamic analysis techniques suffer from some limitation, such as necessity to run malware in virtual machine environment, but this limitation is the trade off for the good results dynamic analysis provides. Thus, we first generate effective model that cannot easily evaded by obfuscation. More accurately, we execute the malware program in a controlled environment and observe its interaction with the operating system.

In summary, our main contribution is to propose a framework based on graph mining approach. System calls are modeled as graphs, representing the program semantic. System calls were monitored because they are the primary interactions of malware with the operating system. Our algorithm infers the system-call graphs from execution traces, and then derives unique graphs that discriminate malware from benign. In other words, our method outperform all previous researches as we know and reached 96.6% detection rate with only 3.4% false positive. In contrast to use of graph mining techniques that are very time-consuming, our method does not take much time to perform. Unfortunately, it is observed that some researches have presented a very high accuracy. A close investigation of the paper reveals that the same data that was used for training were used to evaluate the accuracy. It is a very known error in evaluating the accuracy of a model called overfitting. Results of data mining models should be obtained based on cross validation to ensure evation from overfitting [6].

The rest of paper is organized as follows. Section II describes an overview of the system. Section III encompasses more detail about structure of our system.

Section IV provides experimental results while Section V provides related work. Section VI concludes the paper.

## 2. System Overview

The goal of our system is to effectively and efficiently detect previously unseen and unknown malware. For this our detection method is based on the observation of the execution and monitoring the semantic of malware program in VM (Virtual Machine) environment. To model the program semantics and observe its security behavior, we used system call traces. System calls capture the interaction of program with its environment. Some malware use system calls for activating their malicious payload, so based on this fact; we can understand the malware author intent. In this paper we construct a graph based on system calls trace and our aim is to detect malware programs with high detection rate which outperform lots of previous research. An overview of system can be seen in figure 1.

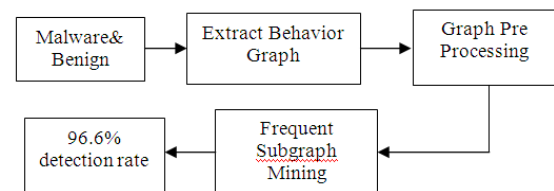


Fig. 1 System Overview

### 2.1 Modeling program semantic

Most of previous research focused on modeling program behavior by specifying permissible sequences of system calls [7, 8]. Malware authors have large degree of freedom in rearranging the code to achieve their goals. For example, it is very easy to reorder independent system calls or to add irrelevant calls. Thus, suspicious activity could not represent as system call sequence that we have observed. Instead a more flexible representation is required. In this paper the representation is based on relationship between system calls and allows rearrangement of system calls. Program semantic is represented as a semantic graph where nodes are (interesting) system calls. An edge is introduced from node x to node y when the return value of system call x is used as an input argument of system call y. Moreover, only a subset of system calls that are essential for

detecting malicious activity (detailed can be found in III) were considered.

## 2.2 Making detection more accurate

While constructing graph we take some considerations into account. That is, edges that its return value is 0x0 or 0x1 are not considered, because these return values means function failure or success. Any other return value is pointer value and is important for constructing the graph. This work makes the graph smaller and it makes any mining process quicker.

## 3. System Details

In this section, more detail about component of the detection system is provided. First our method to diagnose essential system calls for detecting malware is introduced. At the second step we discuss how to characterize program activity via semantic graphs. Then the techniques for extracting graphs automatically from observed traces are discussed. Finally we present our approach to detect graph of previously unknown malicious code.

### 3.1 Essential system call for detecting malicious behavior

Considering all DLL's of windows and all of system calls make graphs very large with lots of unnecessary edge for detection. To address this problem only 6 most important dll (including kernel32.dll, user32.dll, ws\_s32.dll, advapi32.dll, wininet.dll and CreateProcess.dll) were used for malware analysis [9]. In addition, to find subset of system calls in these dll's that are used for malicious activities, data mining was used. Thus, 400 malware and 397 benign applications with all six considered dll's, monitoring all the API's were run, to diagnose which system call are more important for malware detection. Each malware and benign program ran for 3 second. All the system calls that each malware and benign called were collected. Then 10 fold cross validation with random forest classifier [10] were used to measure the accuracy rate of selected system call and the result get 89.5% detection rate. Next we used feature selection techniques to select most discriminative system call. On the other hand based on previous work [9] Malware's operations can be categorized as follows

File access

System information

Networking

Registry access

Processes

System information

It is more important for a malware to gather as much as possible of system information to insure that its software exploit is working. A software exploit is normally related to one specific operating system.

Registry access

In registry a lot of confidential information is stored, like keys or parameters for programs. It furthermore provides a mean to steer the processes that are launched during the machine's boot process. A lot of malware aim to be executed every time when the machine is started.

Processes

A running instance of an executable program is referred to as a process. A process consists of one or more threads, which is an atomic unit when it comes to processor time allocation. All threads that run in the context of a given process share the same address space, security context and environment variables [11].

Networking

The file I/O functions (CreateFile, CloseHandle, ReadFile, ReadFileEx, WriteFile and WriteFileEx) provide the basic interface for opening and closing a communication resource handle and for performing read and write operations. This means that when a process wishes to communicate through a communication device, it can perform a call to CreateFile specifying COM1 or LPT1 or another valid device name, and then write to the returned handle. The process can use the DeviceIoControl-call to send control codes to a device. Several types of malware perform operations against the local network and/or the Internet in order to infect other computers, receive updated malware code or interact with its creators.

We conclude that it is essential to consider all of system calls that are related to above operation for analyzing malware behavior [9]. We also added these system calls to our monitoring file.

### 3.2 Behavior Graphs: specifying program behavior

In general our graph is undirected labeled simple graph. Here is some preliminary concept that is essential for understanding our method [12].

Definition 1(Labeled Graphs) A labeled graph can be represented by a 4-tuple,  $G = (V, E, L, I)$ , where

$V$  is a set of vertices,

$E \subseteq V \times V$  is a set of edges

$L$  is a set of labels,

$l: V \cup E \rightarrow L$ ,  $l$  is a function assigning labels to the vertices and the edges.

This definition can be generalized to include partially labeled graphs if the label set  $L$  includes an empty label.

An example of undirected labeled graph can be shown in figure 2.

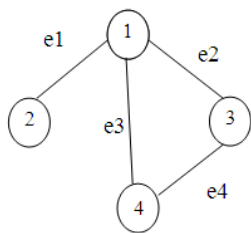


Fig. 2 An undirected labeled graph

**Definition 2 (Subgraph, Induced subgraph)**

A subgraph of a graph  $G$ , is a graph whose vertex set is a subset of that of  $G$ , and whose adjacency relation is a subset of that of  $G$  restricted to this subset.

Given a graph  $G = (V(G), E(G), L(V(G)), L(E(G)))$ , an induced subgraph of  $G$ ,  $G_s = (V(G_s), E(G_s), L(V(G_s)), L(E(G_s)))$ , is a graph satisfying the following conditions.

$$V(G_s) \subset V(G), E(G_s) \subset E(G),$$

$$\forall u, v \in V(G_s), (u, v) \in E(G_s) \Leftrightarrow (u, v) \in E(G).$$

Where  $G_s$  is an induced subgraph of  $G$ , it is denoted as  $G_s \subset G$  [13].

**Definition 3 (Isomorphism, Automorphism, subgraph Isomorphism)** An isomorphism is a bijective function

$$f: V(G) \rightarrow V(G'), \text{ such that } \forall u \in V(G), l_G(u) = l_{G'}(f(u)), \text{ and}$$

$$\forall (u, v) \in E(G), (f(u), f(v)) \in E(G') \text{ and } l_G(u, v) = l_{G'}(f(u), f(v)).$$

An automorphism of  $G$  is an isomorphism from  $G$  to  $G$ . A subgraph isomorphism from  $G$  to  $G'$  is an isomorphism from  $G$  to a subgraph of  $G'$ . If  $f$  is only injective, then  $G$  is monomorphic to  $G'$ .

Induced subgraph isomorphism can be considered as constrained subgraph isomorphism.

**Definition 4 (Frequent Subgraph Mining)** Given a graph dataset,  $GS = \{G_i | i = 0 \dots n\}$ , and a minimum support,  $\text{minSup}$ , let

$$\zeta(g, G) = \begin{cases} 1 & \text{if } g \text{ is isomorphic to a subgraph of } G \\ 0 & \text{if } g \text{ is not isomorphic to any subgraph of } G \end{cases}$$

$$\sigma(g, GS) = \sum_{G_i \in GS} \zeta(g, G_i) \tag{1}$$

$\sigma(g, GS)$  denotes the occurrence frequency of  $g$  in  $GS$ , i.e., the support of  $g$  in  $GS$ . Frequent Subgraph mining is to find every graph,  $g$ , such that  $\sigma(g, GS)$  is greater than or equal to  $\text{minSup}$ . An example of graph mining approach can be show in figure 3.

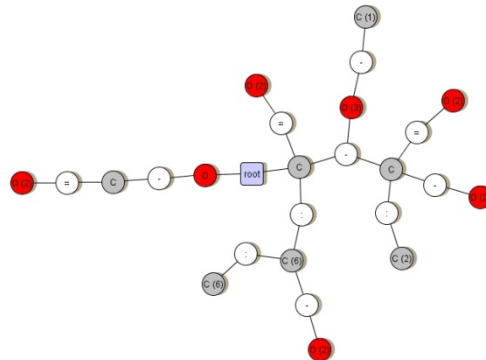


Fig. 3 An example of Graph mining

As a general data structure, graph, specially labeled graph, can be used to model many complicated relation among data. Labels of vertices and edges can represent different attribute of entities and relationship among them. In our setting, label of the nodes are system call



names and label of edges are number of unique values passed between the system calls. Table 1 shows two system calls that have an edge between each other.

Table 1. System calls and their parameters

System call name	Parameters	Return value
CreateFileW	lpFileName:0x00415F2C	0x000025A8
CloseHandle	hObject: 0x000025A8	0x00000001

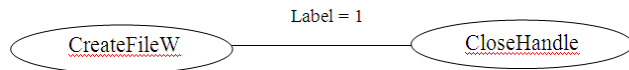


Fig. 4 Making graph based on table 1

If two system calls make an edge with only one memory address then the label of this edge is 1. If two or more unique addresses are used as an input of one API and return value of another, then the numbers of unique addresses are used as the edge label. Figure 4 shows the edge between two system calls with the label of edge and vertexes. Figure 4 is complete example for making edge from CreateFileW to CloseHandle system call based on execution traces. In this example 0x000025A8 is the return value of CreateFileW that is used as an input parameter of CloseHandle, so we draw an edge from CreateFileW to CloseHandle.

### 3.3 Why gSpan was Used

gSpan (graph-based Substructure pattern mining) is used for mining graphs that were generated based on description of previous step. It discovers frequent substructure without candidate generation. gSpan builds a new lexicographic order among graphs, and maps each graph to a unique minimum DFS code as its canonical label. Based on this lexico-graphic order, gSpan adopts the depth-first search strategy to mine frequent connected subgraphs efficiently. This algorithm has very good parallel and scale up properties and can incorporate constraints nicely in graph mining. It can find frequent subgraphs one by one, from small to long ones. Output of this algorithm is as below:

t # id \* support

vertex-edge list

x graph\_id list

where "id" is an integer, the serial number of the pattern, "support" is the absolute frequency of the graph pattern and "graph\_id list" is a list of graphs that contain the

pattern. We used this output for detecting subgraphs that discriminate malicious code from benign one.

### 3.4 Extract Dataset

gSpan was ran with different supports from 0.04 to 0.09. Each frequent subgraph in gSpan is used to be one feature in the final dataset. If one benign or malicious code includes the frequent subgraph, value of that feature is set to 1 otherwise 0 is assigned to the feature.

## 4. Evaluation

404 malware samples and 349 benign samples were collected from [11]. Our system has near optimal detection rate with very low overhead. In this section, system detection capability is presented.

Table 2. Detection Effectiveness of Our System

Name	Number
Constructor	188
Backdoor	162
Exploit	54

### 4.1 System Detection capability

To demonstrate our system detection capability behavior graphs for 3 popular malware families were generated. Table 2 shows an overview of these families and their counts. These malware families were selected because they are very popular according to lists compiled by anti-virus reports [14]. Some of the families use code polymorphism or metamorphism. It makes the detection harder for signature-based scanners. For each malware family more than 50 samples were selected randomly from our database. Specifically samples that did not modify the file system were not used. A single-path dynamic analysis of the samples for 120 second was performed to collect the execution trace. This time is selected because two minutes is generally enough time for most malware to execute its immediate payload, if it has one [15]. While some malware samples do not perform any malicious behavior in this period, these samples usually wait for some external trigger to execute their payload (e.g. network or system environment), and will not perform any behavior if left to execute without further action [15]. Each benign sample also ran for 120 second. The samples were then used for extracting behavior graph. All of the malware and benign graphs used as an input of gSpan to obtain frequent graphs. Because of strong preprocessing step for constructing graph, resulted graphs were very suitable for using graph

mining technique (in terms of size of graph) on the other hand these graphs include all of the information that may be needed for improving detection accuracy. gSpan is used with different support from 0.04 to 0.9 to evaluate different result of this tool. Count of frequent subgraph for each support is in table 3.

Table 3. Detection Effectiveness of Our System

Support	# of frequent subgraph
0.04	4187
0.05	1188
0.06	784
0.07	579
0.08	501
0.09	471

Support 0.9 considered as maximum support because for supports of more than 9 the output includes only graphs with one vertex that is not suitable for our purpose. Each frequent subgraph used as a feature for making final dataset. At the final step, 10-fold cross validation with random forest classifier was used to evaluate the detection rate of the system. Results are shown in Table 4.

Table 4. Detection Effectiveness of Our System

S	Recall	Precision	Fp	F-Measure	ROC Area
0.04	89.9	88.2	10.1	89.1	95.3
0.05	88.7	87.4	11.3	88	95.2
0.06	89.9	87.8	10.1	88.8	95.5
0.07	94.6	96.3	5.4	95.4	98.8
0.08	96.1	98.7	3.9	97.4	98.7
0.09	96.6	98.7	3.4	97.6	99

As shown in table 4, 96.6 percent detection rate with 3.4% false positive was obtained based on 0.09 support. Overall, an average 92.6% detection rate with 7.36% false positive was obtained. Figure 5 illustrates the relationship between detection rate and support, while figure 6 illustrates the relationship between support and false positive.

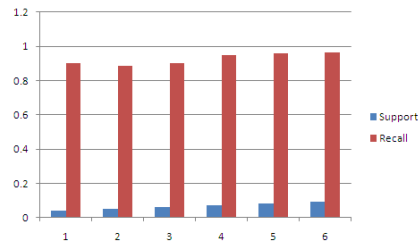


Fig. 5 Support and detection rate relationship

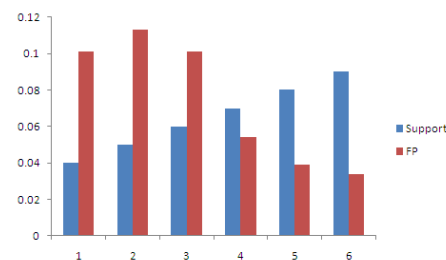


Fig. 6 Support and false positive relationship

## 5. Related Works

Even though behavioral detection seems a recent trend, in antivirus products as well as in virology research, its principles are not really new. In 1986, Cohen [13, 16] already established a basis for behavioral detection within his first formal works. At the other hand, there is a large number of previous works that studies the behavior [17, 18, 19] of different types of malware. Reik et al. proposed classification technique that uses support vector machines to produce class label for unknown malware [20]. Kolbitsch et al. proposed an effective and efficient method for detecting malware behavior at the end host. Their behavior graph was almost like our graph but their graph is more complicated than ours and also they do graph matching as detection method [21]. Egel et al. describe a behavioral specification of browser-based spyware based on taint-tracking [12], and panorama uses whole-system taint analysis in a similar vein to detect more general classes of spyware. Fredrikson and Jha et al. were automated clustering efforts to create initial sample partition for behavior extraction [15]. They demonstrated a technique for producing behavior graphs with 86% detection rate on new, unknown malware, with 0 false positive but they used 912 malware samples and only 49 benign programs for analyzing and these result cannot generalized to other setting. They used input and output parameter for construct graph, this makes graph

larger. We just consider the return value of system call and input value to construct graph. Consequently our graph is simpler. We analyzed 404 malware and 349 benign program and we used random forest classification method to evaluate the detection rate of new, unknown malware. Our result shows 96.6% detection rate with only 3.4% false positive for new, unknown malware.

## 6. Conclusion

Malware detection is a tedious and complicated chore. We propose a method for detecting malicious code from benign based on graph mining techniques that resulted 96.6% detection rate with only 3.4 false positives. Graph is used for representing the behavior of application because graph, especially labeled graph, can be used to model lots of complicated relation between data. At the next step we mined information graph and extracted the most discriminative graphs that separate malware from benign.

## References

- [1] M. Christodorescu, S. Jha, and C. Kruegel. "Mining specification of malicious behavior". In ESEC/FSE. 2007.
- [2] M. Christodorescu, AND S. Jha. "Static Analysis of Executables to Detect Malicious Patterns". In Usenix Security Symposium. 2003.
- [3] M. Christodorescu, S. Jha, S. Seshia, D. Song, AND R. Bryant. "Semantics-Aware Malware Detection". In IEEE Symposium on Security and Privacy. 2005.
- [4] C. Kruegel, W. Robertson, and G. Vigna. "Detecting Kernel-Level Rootkits Through Binary Analysis". In Annual Computer Security Applications Conference (ACSAC). 2004.
- [5] A. Moser, C. Kruegel, and E. Kirda. "Limits of Static Analysis for Malware Detection". In 23rd Annual Computer Security Applications Conference (ACSAC). 2007.
- [6] Y. Ye, D. Wang, T. Li, and D. Ye. An intelligent malware detection system based on association mining. In Journal in Computer Virology, 2008.
- [7] S. Forrest, S. Hofmeyr, A. Somayaji, AND T. Longstaff. "A Sense of Self for Unix Processes". In IEEE Symposium on Security and Privacy. 1996.
- [8] D. Wagner, and D. Dean. "Intrusion Detection via Static Analysis". In IEEE Symposium on Security and Privacy. 2001.
- [9] <http://msdn.microsoft.com/>
- [10] L. Breiman. " Random Forests ". Kluwer Academic Publishers. Manufactured in The Netherlands. 2001.
- [11] A. Sami, B. Yadegari and H. Rahimi, N. Peiravian, S. Hashemi, A. Hamze. "Malware Detection Based on Mining API Calls". SAC'10 March 22-26, 2010, Sierre, Switzerland.
- [12] X. Yan and J. Han. "gSpan: Graph-Based Substructure Pattern Mining". IEEE International Conference. 2002.
- [13] A. Inokuchi, T. Washio, and H. Motoda. "Frequent Substructure from Graph Data". PKDD2000, Sept. 13-16, 2000, Lyon, France.
- [14] F. Cohen. "Computer viruses". Ph.D. thesis, University of South California (1986)
- [15] M. Fredrikson and S. Jha, M. Christodorescu and R. Sailer, And X. Yan. "Synthesizing Near-Optimal Malware Specification from Suspicious Behaviors". IEE Symposium on Security and Privacy. 2010, pp. 45-60.
- [16] F.B. Cohen. "Computer viruses: Theory and experiments". Comput. Secur. 6(1), 22-35 (1987)
- [17] M. Polychronakis, P. Mavrommatis, and N. Provos. "Ghost turns Zombie: Exploring the Life Cycle of Web-based Malware." In Usenix Workshop on Large Scale Exploits and Emergent Threats (LEET). 2008.
- [18] M. , Rajab, J. Zarfoss, F. Monroe, and A. Terzis. "A Multifaceted Approach to Understanding the Botnet Phenomenon". In Internet Measurement Conference(IMC). 2006.
- [19] S. Small, J. mason, F. Monroe, N. Provos, and A. Stubblefield. "To Catch A Predator: A Natural Language Approach for Eliciting Malicious Payloads". In 17th Usenix Security Symposium, 2008.
- [20] K. Rieck, T. Holz, C. Willems, P. Dussel, and P. Laskov, "Learning and classification of malware behavior," in Proceedings of the 5th Conference on Detection of Intrusions and Malware and Vulnerability Assessment (DIMVA'08). Springer, 2008, pp. 108 - 125.
- [21] C. Kolbitsch, P. Milani Comparetti, C. Kruegel, E. Kirda, X. Zhou, and X. Wang. "Effective and Efficient Malware Detection at the End Host". Secure Systems Lab [TU Vienna, Institute Eurecom Sophia Antipolis, UC Santa Barbara]Indiana University at Bloomington. 2009.

**Fatemeh Karbalaie** has obtained her B.S degree in Computer Science in 2007 at Isfahan Payamenoor University. Since 2009, she is a master student of Computer Engineering at Shiraz University. Her research interests include security and data mining.

**Dr. Ashkan Sami** has obtained his B.S. from Virginia Tech; Blacksburg, VA; U.S.A., M.S. from Shiraz University; Iran and Ph.D. from Tohoku University; Japan. He is interested in Data Mining, Software Quality and Security. Ashkan has been a member of technical committee of several international conferences like PAKDD, ADMA, HumanCon, and Future Tech and has more than 40 conference paper and nearly 10 journal papers. He is an associate member of IEEE and was among the founding members of Shiraz University CERT.

**Mansour Ahmadi** has obtained his B.S. in Applied Mathematics from Sistan Baloochestan University; Iran and his M.S in software engineering from Islamic Azad university, Arak. He Worked on malware detection as his M.S. thesis under supervision of Dr. Sami and is currently a researcher in Shiraz University CERT.

# A Multi-agent Approach for Space Occupation Problems

Jamila Boussaa, Mohammed Sadgal and Aziz Elfazziki

Computer Science Department, Faculty of sciences Semailia, University Cadi ayyad  
Marrakesh, BP 2390, Morocco

## Abstract

The occupation of space is a recurring problem in many areas for constraint satisfaction and optimization. The used approaches tend to privilege the optimization or the satisfaction without leading to a general solution. In spite of the success of the few methods of space occupation problems, it can be interesting to consider new ways for resolution, in particular methods resulting from Artificial Intelligence techniques. Because the problem is NP-complex, one possibility of overcoming this complexity is to distribute it across multiple processing units and adopt an appropriate form for decision-making. To construct and evaluate possible solutions for this class of problems, we propose in this paper a general architecture that can accommodate several approaches for resolution through agglomerates of specialized solvers. On this basis, a general model of agent solver is provided. The competences and interactions of agents will be studied and classified according to space occupation problem types. One case is presented here, the resolution by coalition.

**Keywords:** Space Occupation, Constraints, Satisfaction, Optimization, Coalition, Artificial Intelligence, MAS (Multi-Agent System), DSCSP (Distributed Space CSP).

## 1. Introduction

The Space Occupation Problem (SOP) consists of placement of objects in a preset space with respect to a set of imposed constraints; some declared and others not; which appears at the installation time. Such occupation must be *optimal*. The mathematical models dealing with this problem privilege the optimization aspect (occupied space minimization, for example) without explicating or explaining the constraint satisfaction aspect. In fact, the expression of constraints is reduced to an evaluation function that is inadequate in the most cases. Because, the space is simplified in fixed zones (places) and the effort is focused on the assignment cost, the problems with non-numeric expression of constraints can't be resolved. So, the modeling cannot understand the problem as an objects assignment to a set of places.

Many of these problems, undertaken by traditional processing, are confronted with a realistic representation of constraints and objectives. Models as linear programming that require to translate the constraints in equations or statistical models that deal with the problem in term of classification or models based on physical phenomena (Annealing simulated) [1], confirm the difficulty to

generalize the expression of constraints in a numeric aspect. More recently, Thierry Petit and al. [2] study the propagation of side constraints to solve problems. They provide a theoretical and experimental comparison of two main approaches for encoding over-constrained problems with side constraints. Even if their work is oriented constraint programming, the resolution is still problem-dependent.

Generally, constraints are used under an imposed numerical structure that doesn't allow several specialized procedures to cooperate. Indeed, constraints like "*objects using water must be placed as far as possible from any electrical instrument*" or "*the object A must be seen by the object B*" introduce inaccuracy and ambiguity and claim representation techniques and reasoning supported by Artificial intelligence (A.I.) approaches [3,4]. In this sense, the authors in [5] provide an expert approach system in firms.

The problem is NP-complex, thus we can't postulate that only the algorithm performance will be able to overcome the complexity, in spite of the efforts provided in [6]. Some models based on elementary behaviors of reactive agents (i.e ACO and PSO []) had succeeded for certain problems. Whereas other more complex models (cognitive agents) inspired from human behaviors like negotiation, cooperation and game theory still offer approaches to reduce the problem difficulty. These approaches are promising for several reasons such as the expression power to integrate qualitative constraints, need for cooperation between different types of knowledge and modeling by using the "Agent" paradigm [7].

Generally, the SOP is assimilated to a Constraint Satisfaction Problem (CSP). Constraints are treated on a scale of "severities" incorporating the preferences. The priority is to find solutions satisfying "severe" constraints then, order by preferences.

This article presents a multi-agent architecture allowing to express user requests (demands and preferences) easily and naturally and to greet several communities of agents. Each community uses a specific approach for resolution. The goal is double: first, we search to solve CSP with optimization and second, we develop an "infrastructure" able to integrate more than one approach for the resolution. In the following section, we present the State of the art in the field. Section 3 exposes the detailed description of the

suggested approach. A presentation of negotiation and cooperation as a way to resolve the problem by an agents' community based on the coalition will be given in section 4. Section 5 provides an example to illustrate our method. Finally, we conclude by an optimistic note for work to come.

## 2. Related work

### 2.1 Space Occupation Problem (SOP)

Generally, the space occupation problem is expressed by using a CSP or SCSP (Space CSP) [8,10]: Objects are identified by multidimensional variables. A variable would be, for example, a vector of position, orientation and object dimensions. In the most approaches, the constraints are expressed using geometrical relations. But there exist other constraint types representing topological and/or functional nature. According to several authors [8,12], the main difficulty to solve this problem comes from some aspects like the presence of constraints and objectives together. The hardness to optimize antagonistic criteria and to obtain a discrete formulation of the problem leads to NP-Complexity or worse.

### 2.2 Resolution Methods

**Traditional techniques:** To solve this problem, some classical approaches were used. They can be classified in three main categories: The constructive approach which is a top-down approach type according to [8]. The iterative approach tries to improve an occupation of space starting from an earlier one by moving an object or by permuting two objects [10]. The hybrid approach is a coupling, more or less extremely, of two preceding approaches. The basic algorithm is the chronological Backtrack [9]. But this mechanism produces a selectivity problem. Indeed, in the failure case, the algorithm reconsiders the last choice carried out, without worrying to know if this choice has any responsibility in the current failure.

**The distribution aspect, Resolution by MAS/DSCSP:**

A DSCSP uses the traditional definition of a SCSP, by adding the assumption that variables (or constraints) are managed by agents in order to satisfy constraints. Constraints can exist between variables of the same agent (intra-agent constraints) or variables of different agents (inter-agents constraints). Solving the problem is usually seen like carrying out the coherence or the consistency of a multi-agent system. The resolution in a distributed environment allows a parallel processing; therefore time can be saved. But this implies the use of communication mechanisms efficiently in order to ensure the system coherence during the resolution.

In its work, Yokoo [11] has adapted several algorithms: DBA, ABT... The agents are considered responsible for maintaining the environment update. Then, these approaches lead to obtain partial solutions quickly. It might be interesting for dynamic problems that require a great reactivity.

Another approach based on cooperation was imagined by [12] in APO. The agents have a priority and cooperate during mediation meetings. When an agent cannot find consistent value with the more priority agents, it launches a mediation meeting or it changes its value and transmits it to its neighbors. Method ADOPT [13] has for principal application, distributed optimization under constraints. Each constraint is associated with a cost and each agent has to minimize the function 'global objective' (total cost). Since the purpose of the problem is a configuration (assignment of the variables) satisfying the constraints, [14] has imagined reaching this solution by emergence with agents' auto-organization. Thus, the artificial system must fulfill an adequate function. To change function, just change the organization of the system components [26]. Several other authors propose the resolution of certain particular cases by organization and coalition [15,16].

The approaches above improve conventional algorithms by introducing the distribution and often in the guise of parallelism while admitting certain assumptions such as communication by messages according to Yokoo [11]. But, the majority of these systems encounter a problem on the communication level where it is necessary to manage a great number of messages generated by the agents. The idea of the solution emergence is interesting but the research of the specific adequate functions is very difficult. So we are looking for our approach to solve the problem by a collective decision by all agents using coalition and deliberation mechanisms.

## 3. Suggested approach

### 3.1 General architecture

We present a MAS architecture allowing to adapt resolutions by using several types of agents, from a simple reactive ("reflex") agent to a cognitive one more complex. The recourse to a multi-Community architecture (figure 1) of contextual agents depending on the CSP categories is justified by several arguments: 1) the resolution is perceived as the effort of several units, each one contributes by a partial or total solution. 2) The interpretation of the problem expressed by the user belongs to the resolution. The presence of the highly cognitive agents is useful in the clarification of the user requests that are often too general. 3) Adapting the algorithm to the problem (or the reverse) influences the solution's quality. 4) The parallelism question, as reconsidered under the Distributed A.I. MAS

Paradigm, offers new possibilities to converge towards solutions adapted via: competition, cooperation, negotiation, organization...

Our long-term goal is to offer a system model able to receive several communities of agents (solvers [17]), each one is specialized in the resolution of a class of problems. The community has its own behaviors and has its own methods of resolution. The Interface agent (Supervisor) deals with the interpretation of the initial problem and the contexts for the specialized communities. The solutions suggested by a community can be retained by the supervisor according to some evaluation criteria. A definite decision will be concerted with the user (figure 1).

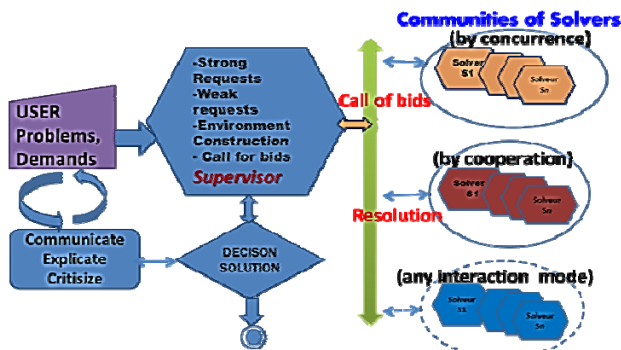


Fig. 1. General architecture of the MAS

### a. Types of agents

**User Agent (UA):** The User agent represents the user. He is the initiator of the problem in form of descriptions concerning space occupation, objects to be placed and the demands (constraints and objective). He communicates with the Supervisor agent through an interface.

**Supervisor Agent (SA):** The Supervisor agent is a “mediator” between the communities of the solvers and the User agent. It represents descriptions and demands into a Base of facts and rules in predicate logic. By using rules of transformation, the requests are converted into “severe” constraints which must be respected obligatorily and into preferences that are hopes to be carried out as well as possible. The SA has the heavy task for selecting the community that would be “able” to solve the problem by using a set of problem categories and task announcements. So this agent has a diagram even an ontology enabling it to classify the problem (placement, cutting, routing...) and to suggest the community of suitable solvers.

**The Community of agents:** A community consists of homogeneous agents equipped with competences and are specialized in the resolution of one or more classes of problems according to a resolution model (competition, cooperation..., (see section 4).

### b. Environments

The choice of the environment and its properties is closely related to the CSP type and to the singularity of the community. In general, within the framework of a closed system, it is possible to determine neighbors for each agent. One way to formulate is to fix them initially since their creation. The choices, made during the construction of the neighbors, can orient a model in any direction.

### c. Interaction of agents within a community

How agents interact and how they are organized make them to coordinate themselves, to cooperate or to negotiate. Coordination is an essential point, especially with respect to process implementation of the multi-agent models, to determine which does what and when is a non-trivial problem that can have infinity of solutions. Each one of them can appreciably modify results obtained from simulations [18]. Figure 2 recapitulates what will be integrated in the system.

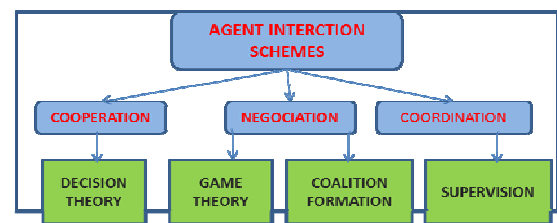


Fig. 2. Diagrams of interaction

The decision theory, where an agent tries to maximize a criterion (called utility), is rather close to the game theory. The difference between decision and game theories is that the game theory takes into account the current situation and also the future choices of agents. The coalition formation is another approach used in interactions between agents. It acts, for agents confronted with a request, to make the individual compromise in order to reach a consensus for all parts (ideal case). Then, difficulty will define the communication protocol in an adequate way. The protocol must make agents exchange their current choices and modify them until realization of consensus (see section 4).

### 3.2 Basic concepts common to agent communities

In order to provide a general structure for the SOP resolution, we present here a description based on the following definitions:

A placement space of two or three dimensions in which geometrical objects with possible functional and topological characteristics will be placed. The installation is governed by all constraints using characteristics of objects and space. The goal is to occupy space by installing all objects, satisfying constraints and carrying out objectives as well as possible.

In the distributed version of a CSP, authors traditionally distribute variables or constraints on agents. Each agent is

given the responsibility to solve its problem locally while contributing to the global resolution. We consider that each agent deals with one object to place in the space (figure 3).

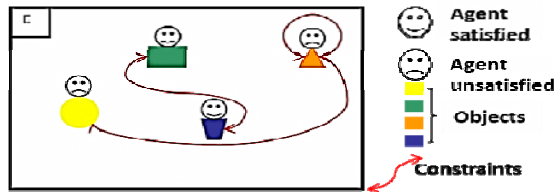


Fig. 3. Occupation of space by agents (objects)

### 3.2.1 Notations and definitions

In our model, the physical world is abstracted in States, Actions and State Transitions caused by actions. A State is the assignment of all objects (agents) to places in the occupation space. A State Transition relates to the passage from one state to another. Lastly, an action is the way by which a transition will be realized. The resolution will be seen like series of transitions to pass from an initial State to a final State (solution). Let us consider the following definitions:

- E:** space in **2D** or **3D** (e.g rectangle  $(x_0, y_0, x_d, y_d)$ )
- A** =  $\{A_1, \dots, A_n\}$  set of agents (agent represents an object)
- S** =  $\{S_0, \dots, S_m\}$  set of States
- C** =  $\{c_1, \dots, c_p\}$  set of constraints
- B** =  $\{b_1, \dots, b_q\}$  set of objectives
- AC** =  $\{a_1, \dots, a_r\}$  set of actions of agents
- An action  $a_j$  is regarded as the joint action of all the agents:  $a_j = (a_{j1}, \dots, a_{jn}, \dots, a_{jn})$  where  $a_{ji}$  is the  $i^{th}$  agent action.
- P** =  $\{p_1, \dots, p_t\}$  set of plans
- A plan  $p_i$  is a set of joint actions:  $p_i = \{a_0, a_1, \dots, a_k\}$
- Actions:** An *individual* action “a” of an agent is a change of its place in space E. This change can be performed using combination of geometrical operators like translation and rotation (in certain cases of design, the action can also be a change of object dimensions).
- For example (figure 4):  
 $a = (\delta t_u, \delta t_v, r_w) \in R^3$ , if  $P_i$  = position occupied by  $A_i$  in E,  
 $P_i = (x_i, y_i, t_i)$ : location  $(x_i, y_i) \in E$  and orientation  $t_i$  of the local reference (object reference).  
 $a(P_i) = P'_i$ : Change of place and orientation of  $A_i$  by application of action  $a = (t_u, t_v, r_w)$ .  
 Then  $P'_i = (x'_i, y'_i, t'_i)$  where  $x'_i = x + \delta t_u$ ,  $y'_i = y_i + \delta t_v$  and  $t'_i = t_i + \delta r_w$   
 $a_0 = (0, 0, 0)$  is the action identity: the agent does not move.

**Place:** A place for an agent is defined by a geometrical position  $P_i$  (Coordinates, Orientation) and object Dimensions. We note this place  $P_i = (P_i, Dg_i)$ , where  $Dg_i$  are geometrical object dimensions (e.g. length and width for a rectangular object in **2D**).

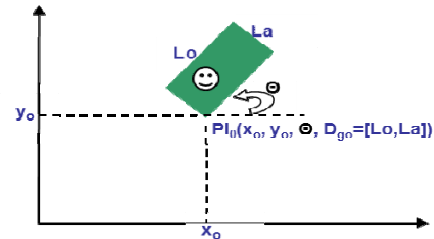


Fig. 4. An object in a referential related to space E

**State:** We define a State  $S_k$  like a triplet  $\langle PL_k, C_k, B_k \rangle$ :  
 $PL_k = \{P_{k1}, \dots, P_{ki}, \dots, P_{kn}\}$  is all places occupied by agents in E ( $P_{ki}$  is the place of agent  $A_i$  in  $S_k$ ).  
 $C_k$  is the subset of satisfied constraints in the State  $S_k$   
 $B_k$  is the subset of objectives achieved in the State  $S_k$

**States' Transition:** A transition on States is defined by the  $(S_k, S_l)$  pair, and denoted  $S_k \rightarrow S_l$ . State  $S_l$  is regarded as a consequence of the action causing the  $S_k \rightarrow S_l$  transition on  $S_k$ . We note  $(S_k \rightarrow S_l) / a_j$  to indicate the transition and its cause  $a_j$ .

### 3.2.2 Admissibility and optimality

**Admissibility:** C is the set of “severe” constraints in the SOP with cardinal  $(C) = |C| = p$ . A constraint is seen like a relation (arcs on figure 3) between one or more agents. A State  $S = \langle PL_S, C_S, B_S \rangle$  is known as **admissible** if and only if  $C_S = C$  (situation where all the constraints are satisfied).  
**Note:** The relation “same set of satisfied constraints” is a relation of equivalence on the set of States  $\mathcal{S}$ . Thereby, the **admissible** States constitute a class of equivalence by: “they have C like set of satisfied constraints”. Let  $\mathcal{E}_a$  this class,  $\mathcal{E}_a = \{S \in \mathcal{S} \mid S = \langle PL_S, C, B_S \rangle\}$ .

**Agent level satisfaction:** If  $C^{A_i}$  indicates the set of constraints for agent  $A_i$ , then:  $\cup_i C^{A_i} = C$  with  $A_i \in A$ . State  $S = \langle PL_S, C_S, B_S \rangle$  satisfying the agent  $A_i$  is said **Ai-admissible** if and only if  $C^{A_i} \subset C_S$ . Generally,  $S = \langle PL_S, C_S, B_S \rangle$  satisfies a group of agents  $G_A$  if  $\cup_i C^{A_i} = C$  with  $A_i \in G_A$ .  
**Notice** (based on individual satisfactions of the agents): For any State S if  $\forall i, C^{A_i} \subset C_S$  then S is admissible.

**Optimality:** The goal is to find an admissible State optimizing the objectives (preferences): These preferences can be regarded as less severe constraints, but it will pose a difficult problem: the optimization. Most of the applications propose optimization by seeking more adequate models. On our side, we suppose the existence of an evaluation function

measuring a realization degree for objectives. Thus, our problem is to define an evaluation function based on each agent's appreciation.

**Appreciation:** An agent can evaluate any State  $S$  according to the satisfaction of its own constraints, the remainder of constraints and objectives. Then, evaluation is defined by the function  $ju: \mathcal{A} \times \mathcal{S} \rightarrow [0,1]$ .

$ju(A_i, S)$  expresses a satisfaction degree (constraints and preferences) by agent  $A_i$  for State  $S$ . This value is calculated by the agent  $A_i$  taking into account its position in  $PLs$ , subset of satisfied constraints and objectives carried out, locally and globally.

### 3.2.3 Elements of resolution

To solve the problem, it is primordial to have  $\mathcal{Ea}$  (set of admissible solutions). In this case, procedures must determine the optimal solution in  $\mathcal{Ea}$ .

Given the problem complexity, we emit two assumptions:

- To prove the existence of  $\mathcal{Ea}$  is not necessary to begin the resolution.
- The optimal solution search is evaluated in terms of convergence on  $\mathcal{Ea}$  (objectives are carried out on acceptable States rather admissible States)

#### The search for an acceptable State:

An  $A_i$ - $k$ -admissible State is a State where  $k$  constraints are not satisfied ( $k$ : dissatisfaction degree) for the agent  $A_i$ .

An  $A_i$ -admissible State is an  $A_i$ -0-admissible State (all  $A_i$  constraints are satisfied:  $k=0$ ).

A State is called **acceptable** if it is admitted admissible even if  $k \neq 0$ .

We will say that  $A_i$  and  $A_j$  are **neighbors** if there is at least a constraint between  $A_i$  and  $A_j$ .

Let  $\mathcal{A}_{Gi}$  the  $A_i$  neighbors' set.

We present a **Pseudo-algorithm** to filter the  $A_i$ - $k$ -admissible States with the smallest degree  $k$ :

```

***** for each agent Ai and Sa : actual State *****/
Search_Ai_k_admissibles(Ai, Sa) {
    kai      : current dissatisfaction degree in Sa for the agent Ai
    k=0      : initialization of degree dissatisfaction at the begin of this search
    ESi = ∅  : ESi : list to contain possible A-k-admissible States
    Pi = ∅   : Pi : lists of plans to reach the A-k-admissible States.
    1-find Zik(Sa): zone of possible places with k unsatisfied constraints of Ai.
      If Zik(Sa) != ∅ do:
        - For each Pij (possible place) for Ai in Zik(Sa) do:
          determine the pijk plan and Sijk such as (Sa → Sijk)/pijk
          with Sijk = <PLi, Ci, Bi>
          ESi = ESi ∪ {Sijk}; Pi = Pi ∪ {pijk}
        - End For
      If ESi != ∅ go to 2 endif
    Else
      k = k+1
      if k ≥ kai to go to 2
      else go to 1
    endif
  Endif

  2-If ESi = ∅, ESi = {Sa}, k=kai
  else
    consider Sijk, and record:
    the degree of dissatisfaction k.
    the pijk plans to reach Sijk from Sa.
    the evaluation of Sijk by ju.
  endif
}
    
```

#### Notes:

- All agents are executed in a parallel way; the algorithm provides the possible places when the domain is discrete. To avoid an intensive CPU time, the agent may save only the envelope of the  $Z_{ik}(Sa)$  zone and determine places by applying associated plans in time of need.
- The Search Algorithm can also cover neighboring: Every acceptable State  $S \in ES_i$  for  $A_i$  must be validated by its neighbors ( $\mathcal{A}_{Gi}$ )

**Ai-Optimal State:** The first task of each agent is the generation of plans leading, at least, with  $A_i$ - $k$ -admissible solutions. But with objectives, the agent will seek to propose the "best plan" within the meaning of the appreciation  $ju$ .

The **Ai-optimal** State is  $S_{opi}$  such as:

$$S_{opi} = \arg(\text{Max}(ju(A_i, S_i))) \text{ with } S_i \in ES_i$$

(Ai-k-adm)

We also note  $u_{ji} = ju(A_j, S_i)$  for very  $S_i \in ES_i$ : the  $S_i$  evaluation by the  $A_j$  agent

## 4. Resolution Techniques

**Multi-criteria traditional techniques:** With a utility function, the problem can be solved in a centralized way according to mathematical models and using the multi-criterion techniques [19]. These techniques establish an order on some offered possibilities and the decision maker (human) has to arbitrate. Thus, in our case, the agents will be able to replace the decision makers and the problem will be multi-criterion and multi-decision makers.



Indeed, the numerical measurement of the agent utility is already a strong assumption compared to the simple classification of the available choices. The comparison of the utility of two individuals is even difficult.

Why would a plan, appreciated with 0.8 by an agent and 0.5 by another, be preferred with that respectively appreciated 0.4 and 0.9? It is difficult to represent the importance of a decision maker by a weight in order to respect the general structure of these techniques. The disadvantage is that the aggregation procedures and criteria transformation into constraints are delicate to carry out especially when certain decision makers and interlocutors have no scientific culture.

**Negotiation / Cooperation:** In order to build a model of negotiation by a multi-agent system, some elements must be defined. In [20], authors identify three components as being most fundamental:

- 1- *Negotiation Protocol:* It is a set of rules managing the interaction.
- 2- *Negotiation Object:* It consists of attributes on which agents wish to find an agreement
- 3- *Decision Strategy or model of the agents:* It is a reasoning process used by agents, in agreement with negotiation protocol, to achieve their goals. In literature, there exist three great approaches of negotiation in the multi-agent systems [21] based on:
  - *The game theory:* it studies the behavior of a “rational” agent confronted with one (many) adversary during a game, in order to find an optimal strategy maximizing its own utility. Several protocols were studied [22]. The game theory also uses strong concepts of convergence based on the Nash equilibrium or the Pareto's optimum.
  - *Heuristics:* The lack of resources and time does not make it possible to elaborate a better policy by analysis with game theory. In order to mitigate these limits, the heuristic approaches try to reach acceptable approximations with the theoretical optimal results found by the game theory [23].
  - *Argumentation:* The argumentation is an adequate model to represent the internal reasoning of an agent, and it is based on the construction of arguments. It considers the model interactions of multi-agent in form of dialogs [24].

#### 4.2 A proposition: resolution by forming coalition among agents

In a community of cooperative agents, an agent uses a coalition when it is unable to satisfy all its constraints, i.e. when it cannot find **A<sub>i</sub>-0-admissible** States. Here's an example of resolution based on the negotiation using an analysis by the game theory, among different other models of negotiation (cited above), that will be integrated in the system. As it is difficult to incorporate utilities of agents, an agent will seek an accepted plan (or plans) by all its

neighbors. This plan must produce a State which is “better” or at least equivalent of the Actual State. The solution can be achieved with a **Pareto's optimum** (in the game theory): an agent cannot increase its utility if at least another utility agent is decreased. The protocol of negotiation is based on this principle [25]. The agent initializing the negotiation seeks the plans which it prefers. It transfers them by grouping and ordering to an agent close to its choice. The agent receiving plans, filters those “better” than the actual state, reorders them according to its utility, and sends them with the same procedure to an agent of its choice. The last agent selects the most interesting plan (or plans) that will constitute the Pareto's optimum.

##### 4.2.1 Definitions

**Coalition:** A coalition is a subset  $S \subseteq A = \{A_1, \dots, A_n\}$ , where **A** is the agents' set ( $2^n - 1$  coalitions are possible).

In DCSP and SOP contexts, the resolution by coalition was applied to several problems such as the task allocation, resource allocation... But the CSP differs from those problems that define coalition structure a priori. Indeed in CSP there is a strong dependence, the agent choices are not fixed and change permanently by the others' choices.

Then, agents form a coalition around one or several constraints to find a solution. Several neighbor form a coalition and provide concerted action plans to satisfy the joint constraints as well as possible.

**Set of coalitions:** a set representing a solution to the problem of coalitions' formation. Agents form coalitions to satisfy constraints with objectives. It is about the set of plans which provides a **State solution** in our case. Then we denote by **Group** a set of coalitions' sets.

**Context:** the parameters taken into account in the problem (must be stable during the negotiation).

**Utility function:** the utility function can be ordinal or cardinal. The cardinal associates a utility with a set of coalitions and a given context. The ordinal permits to compare two sets in a given context. In this case, to measure the utility of a **State** means to compare it with a reference **State** (see section 4.4).

**Reference State:** The agents must know if they accept “States solutions”, so it is necessary that they can compare a State with what they are able to obtain during the negotiation. This minimum is the reference State.

##### 4.2.2 Negotiation Algorithm

Each negotiation proceeds in three phases:

**Phase 1:** Initialization of the negotiation and transfer of constraints. The initiating agent informs all the others that it begins a new negotiation. Any agent which will want to begin another from them will have to await the end of the actual negotiation. The initiating agent calculates all the possible coalitions. It gathers them in group of solution sets

and sends it to itself and/or to the agent which must begin the negotiation.

**Phase 2:** Negotiation

When an agent receives a group of sets: it preferably classifies by order (with its utility) the received sets in homogeneous groups. It classifies only the sets at least equivalent, with its reference State. Then, groups are sent to the following agent by a decreasing order.

If all agents already took part in the negotiation (the agent is thus the last). **So** at least one of the sets received is acceptable, it considers the best set. This set is **Pareto's optimum**.

**Phase 3:** Transmission of the solution

Once the last agent identified a Pareto's optimum, it transmits this set to all agents which accept it as a solution of the negotiation.

**Resolution Steps for SOP:** Due to the problem complexity, each agent works initially to satisfy its constraints. For those not satisfied, it will form a coalition with the agents implied in these same constraints. The problem looks like a “repetitive game”, the solution of the problem can be obtained using several negotiation rounds.

1. A starting State **So** is given:  
 Several heuristic can be used here, for example:
  - **So** is the first space occupation without constraints.
  - Since there is dependence, one can proceed by a sequential occupation: an order is established and each agent will seek the plan, by regarding the placed agents.
2. With the reference State (**So** at the beginning), the algorithm (section 3.3.1) is carried out. Any agent **A<sub>i</sub>** which is not able to propose an **Ai-0-admissible** State will seek to form a coalition with its neighbors.
3. A coalition solves the problem by negotiation according to algorithm 4.2:  
 Plans provided into 2. are considered and evaluated by each member of the coalition, the initiating agent of the negotiation orders plans by groups according to its utility and sends them to an agent of its choice, this one retains only those that provide “better” States and so on, to the last agent. Plans retained by the last will constitute the solution. If the State obtained is **Ai-0-acceptable** for each **A<sub>i</sub>**, then the negotiation is finished. If not the actual State will be regarded as reference State and it begins again since 2.
4. Without improvement and at the end of a number predefined of negotiation rounds an agent decides to stop the formation of coalitions.

**4.2.3 Determination of agent utility function (SOP)**

We specify here how to calculate utility value **u<sub>ij</sub>** by **A<sub>j</sub>** agent for a plan **p<sub>i</sub>**, suggested by the agent **A<sub>i</sub>**.

Although it is difficult to model this evaluation using a quantitative function, it is necessary to use certain indices: (i) Satisfaction rate of constraints relating to the agent, (ii) Potential utilization ratio, (iii) Total satisfaction rate and (iv) Satisfaction neighborhood rate.

**a. Definition of rates**

Let:

- S:** actual state,
  - E:** Occupation Space,
  - C:** set of constraints,
  - C<sup>A<sub>j</sub></sup>:** set of constraints of agent **A<sub>j</sub>**,
  - Z<sub>S</sub><sup>A<sub>j</sub></sup>:** satisfaction zone of **A<sub>j</sub>** in **S**,
  - C<sup>J<sub>si</sub></sup>:** satisfied set constraints of **A<sub>j</sub>** in **S<sub>i</sub>**
- a<sub>i</sub>** is action of **A<sub>i</sub>** with **a<sub>i</sub>(S) = S<sub>i</sub>**; (or **(S↔S<sub>i</sub>)/a<sub>i</sub>**)

We call:

- Relative satisfaction rate in **S<sub>i</sub>**: **r<sub>ij</sub> = 1 - |C<sup>J<sub>si</sub></sup>|/|C<sup>A<sub>j</sub></sup>|**
- Potential utilization ratio in **S<sub>i</sub>**: **z<sub>ij</sub> = |Z<sub>S</sub><sup>A<sub>j</sub></sup>|/|E|**
- Total satisfaction rate in **S<sub>i</sub>**: **g<sub>i</sub> = 1 - |C<sub>Si</sub>|/|C|**
- Satisfaction neighborhood rate: **n<sub>i</sub> = 1 - |C<sup>A<sub>gi</sub></sup><sub>Si</sub>|/|C|**; **Ag<sub>j</sub>** is the set of **A<sub>j</sub>** neighbors

**b. Utility**

We can model the utility **u<sub>ij</sub>** that is an **A<sub>j</sub>** judgment on the **A<sub>i</sub>** action by using a linear combination as follows:

**u<sub>ij</sub> = w<sub>1j</sub>\*r<sub>ij</sub> + w<sub>2j</sub>\*z<sub>ij</sub> + w<sub>3j</sub>\*g<sub>i</sub>** (we use **g<sub>i</sub>** when all agents are linked by constraints, otherwise **n<sub>i</sub>**)

The term: **w<sub>1j</sub>\*r<sub>ij</sub> + w<sub>2j</sub>\*z<sub>ij</sub>** expresses the personal interest of the **A<sub>j</sub>** agent

The term: **w<sub>3j</sub>\*g<sub>i</sub>** expresses the global interest

Let: **u<sup>p</sup><sub>ij</sub> = r<sub>ij</sub> + z<sub>ij</sub>** and **u<sup>g</sup><sub>ij</sub> = g<sub>i</sub>**

If **w<sub>1j</sub> = w<sub>2j</sub> = α<sub>j</sub>** and **w<sub>3j</sub> = β<sub>j</sub>** then

**u<sub>ij</sub> = α<sub>j</sub>\*u<sup>p</sup><sub>ij</sub> + β<sub>j</sub>\*u<sup>g</sup><sub>ij</sub>**

With **β<sub>j</sub> = 1 - α<sub>j</sub>**: the more one privileges the personal interest, the more it ignores the global interest and vice-versa (**α<sub>j</sub> ∈ [0,1]**).

Finally: **u<sub>ij</sub> = α<sub>j</sub>\*u<sup>p</sup><sub>ij</sub> + (1 - α<sub>j</sub>)\*u<sup>g</sup><sub>ij</sub>**

Each **A<sub>j</sub>** agent adopts its own strategy (choice of **α<sub>j</sub>**) to calculate its preference (i.e. **α<sub>j</sub> = 1/2** is a neutral strategy)

**4.2.4 Discussion**

In our open architecture, we can use any form of cooperation. So, we use a *Generic Cooperation-based Method definition* [26] that is held on: (i) Cooperation can be viewed as a generic concept manipulated by problem solvers, (ii) It transcends to all the CSP methods, (iii) Taking inspiration from biological and socio-economic notions of cooperation and (iv) An agent alone is unable to find the global solution and it has to interact locally with its neighbors in order to find its current actions and to be able to reach its individual goals and help its neighbors

Thus, it can produce some categories of Cooperation-based algorithms as the Population-based approaches inspired by evolution and the behavior of insects, birds...

Their principles are: (i) A population is a set of individuals (agents), (ii) Each agent is able to find a solution to the problem and (iii) An agent knows the whole set of variables that define the problem. Agents coordinate to find a solution.

The common problem is how to coordinate several concurrent searches to efficiently find a good solution?

Several methods are essentially used in optimization problems: Evolutionary algorithms, genetic algorithms (GA), Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO) [16].

In ACO, The pheromone deposited by ants gives relevant information about the region of the search space and modifies later the behavior of the other ants.

In PSO, Particles are influenced by the velocity and position of the local and global bests: cooperative information exchange allowing efficient exploration phase.

The fitness function of GA determines at a time the better individuals which will share their genes with other members of the population to produce new relevant offspring.

The essential difference with the coalition resolution: the cooperation is based on negotiations using game theory. Agent has a pseudo-global vision (must know its neighbors) and not a local vision. It offers, accepts or rejects solutions in concert with its neighbors. The utility function takes into account local and global aspects (see 4.2.3).

### 5. An illustrative example

To simplify, we consider 4-queens problem. There is no optimization. Only rigid constraints to satisfy, namely: "neither jointed line nor diagonal with the occupied squares". We will restrict the definition of the utility as follows:  $u_i^p$ : number of satisfied constraints for one agent  $A_i$ , and  $u_i^g$ : number of satisfied constraints for all agents. Each  $A_i$  represents one queen  $i$ . All agents are neighbors.

With  $\alpha_i=1/2$ , then, for  $A_i$  we have:  $u_i = 1/2(u_i^p + u_i^g)$ . Let us consider for example, the  $S_0 = (1,1,1,1)$  positions of the 4 agents (line occupied by agent  $A_i$  in its column).

Let us notice that for two agents, the maximum number of non-satisfied constraints is 1. Thus, for an agent, the maximum number of non-satisfied constraints is 3.

If  $A_1$  cannot satisfy all its constraints (whatever its position), then it forms a coalition with its neighbors ( $A_2, A_3$  and  $A_4$ ).  $S_0$  is the reference State at the beginning.

$$S_0 =$$

We note:

$$- U_p = (u_1^p, u_2^p, u_3^p, u_4^p),$$

$$- U = (u_1, u_2, u_3, u_4) \text{ with } u_i = 1/2(u_i^p + u_i^g) \text{ and } u_i^g = u_1^p + u_2^p + u_3^p + u_4^p$$

At  $S_0$ , utilities of agents are identical:

$$U(S_0) = (u_1(S_0), u_2(S_0), u_3(S_0), u_4(S_0)) = (0,0,0,0)$$

The plans (thus States obtained by these plans) suggested with the first round:

**States proposed by  $A_1$ :**

$$S_{11} = (2,1,1,1) \rightarrow U_p(S_{11}) = (2,0,1,1) \text{ from where } u_1 = 1/2(2 + (2+0+1+1))=3, u_2=2, u_3=u_4=5/2, \text{ thus } U(S_{11})=(3,2,5/2,5/2)$$

$$S_{21} = (3,1,1,1) \rightarrow U_p(S_{21}) = (2,1,0,1) \rightarrow U(S_{21})=(3, 5/2,2,5/2)$$

$$S_{31} = (4,1,1,1) \rightarrow U_p(S_{31}) = (2,1,1,0) \rightarrow U(S_{31})=(3,5/2,5/2,2)$$

**States proposed by  $A_2$ :**

$$S_{12} = (1,2,1,1) \rightarrow U_p(S_{12}) = (0,1,0,1) \rightarrow U(S_{12})=(1,3/2,1,3/2)$$

$$S_{22} = (1, 3,1,1) \rightarrow U_p(S_{22}) = (1,2,1,0) \rightarrow U(S_{22})=(5/2,3,5/2,2)$$

$$S_{32} = (1,4,1,1) \rightarrow U_p(S_{32}) = (1,3,1,1) \rightarrow U(S_{32})=(7/2,9/2,7/2,7/2)$$

**States proposed by  $A_3$ :**

$$S_{13} = (1,1,2,1) \rightarrow U_p(S_{13}) = (1,0,1,0) \rightarrow U(S_{13})=(3/2,1,3/2,1)$$

$$S_{23} = (1, 1,3,1) \rightarrow U_p(S_{23}) = (0,1,2,1) \rightarrow U(S_{23})=(2,5/2,3,5/2)$$

$$S_{33} = (1,1,4,1) \rightarrow U_p(S_{33}) = (1,1,3,1) \rightarrow U(S_{33})=(7/2,7/2,9/2,7/2)$$

**States proposed by  $A_4$ :**

$$S_{14} = (1,1,1,2) \rightarrow U_p(S_{14}) = (1,1,0,2) \rightarrow U(S_{14})=(5/2,5/2,2,3)$$

$$S_{24} = (1, 1,1,3) \rightarrow U_p(S_{24}) = (1,0,1,2) \rightarrow U(S_{24})=(5/2,2,5/2,3)$$

$$S_{34} = (1,1,1,4) \rightarrow U_p(S_{34}) = (0,1,1,2) \rightarrow U(S_{34})=(2,5/2,5/2,3)$$

All these  $S_{ij}$  solutions can be retained by any  $A_j$  agent because  $u_j(S_{ij}) > u_j(S_0)$

$A_1$  initiates negotiation; then forms 6 groups of plans ordered (decreasing order) according to its utility  $u_1$ :

Groups in decreasing order:  $G_1 = (S_{32}, S_{33})$ ;  $G_2 = (S_{11}, S_{12}, S_{13})$ ;  $G_3 = (S_{22}, S_{14}, S_{24})$ ;  $G_4 = (S_{23}, S_{34})$ ;  $G_5 = (S_{13})$  and  $G_6 = (S_{12})$ . These groups are sent in this order to  $A_2$

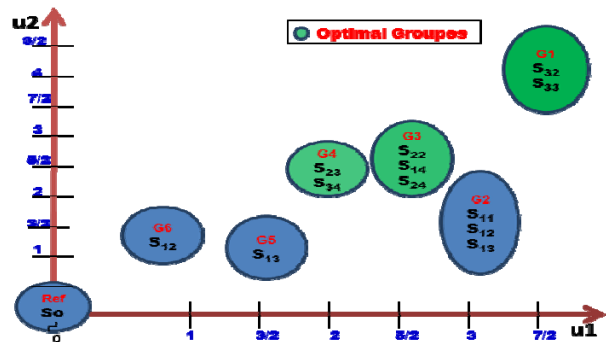


Fig. 5. Graph representing the utilities of the agents  $A_1$  and  $A_2$

$G_1$  is sent at first to  $A_2$  agent (chosen by  $A_1$ , heuristics can be used),  $A_2$  evaluates  $G_1$  according to its utility here:  $G_1 = (S_{32}, S_{33})$ ,  $u_2(G_1) = (9/2, 7/2)$ , which would be better than:  $u_1(G_1) = (7/2, 7/2)$

$A_2$  will then keep  $G_1$  in its entirety and sends it to  $A_3$  (according to its choice), this one will keep  $G_1$  because  $u_3(G_1) = (7/2, 9/2)$ , and sends it finally to  $A_4$ ,  $u_4(G_1) = (7/2, 7/2)$  and decides to keep it because it is better than  $u_4(S_0)=0$ .

X	X	X	X

The satisfaction solution is  $S_f$  with  $U(S_f) = (15/2, 15/2, 15/2, 15/2)$  and is not reached yet, the agents will decide on  $S_{32}$  or  $S_{33}$  to continue, let us suppose that  $S_{32}$ . (because there is no objectives for comparison)  
 The reference State then is changed:  $S_0 \leftarrow S_{32}$ , with:  $u_1(S_0) = 7/2$ ,  $u_2(S_0) = 9/2$ ,  $u_3(S_0) = 7/2$  and  $u_4(S_0) = 7/2$ ,  
 Agents proceed to another round ( $2^{nd}$ ) of negotiation. We have two propositions:  
 $S_{11}^2 = (2, 4, 1, 1) \rightarrow Up(S_{11}^2) = (3, 3, 2, 2) \rightarrow U(S_{11}^2) = (13/2, 13/2, 6, 6)$   
 $S_{24}^2 = (1, 4, 1, 3) \rightarrow Up(S_{24}^2) = (2, 3, 2, 3) \rightarrow U(S_{24}^2) = (6, 13/2, 6, 13/2)$   
 $S_{11}^2$ , is chosen according to the same process,  $S_0 \leftarrow S_{11}^2$ , with  $U(S_0) = (13/2, 13/2, 6, 6)$   
 In the last round ( $3^{th}$ ): the State  $S_{14}^3 = (2, 4, 1, 3)$  proposed by  $A_4$  is accepted by all other agents:  
 $U(S_{14}^3) = (15/2, 15/2, 15/2, 15/2)$ , thus  $S_{14}^3$  is a solution (satisfying all the constraints) and end of negotiation.

$S_{14}^3 =$

		X	
X			
			X
	X		

This problem was used by all CSP algorithms for tests like:  
 -The *nogoods* (conflictual configurations) and potential solutions communicated by agents to their neighborhood in ABT or AWCS cooperatively [11].  
 -The heuristic min-conflict used ERA is a means to represent the fact that agents cooperatively act by minimizing the negative impact of their actions  
 - Population-based approaches (ACO, PSO, GA...)

In our example, we want just to show that CSP solution can be obtained under game theory as a pareto-optimal or nash-equilibrium. By comparison, we quote the model ERA (Environment, Reactive rules and Agents) [27] to have an idea of utility function that is reduced to the number of constraint violations:  
 In solving a CSP with ERA method, each agent represents a variable and its position corresponds to a value assignment for the variable. The environment for the whole multi-agent system contains all the possible domain values for the problem, and at the same time, it also records the *violation numbers* for all the positions. An agent can move within its row, which represents its domain. Three reactive behaviors (rules) were introduced: *better-move*, *least-move*, and *random-move*. The move of an agent will affect the violation numbers of other rows in the environment.

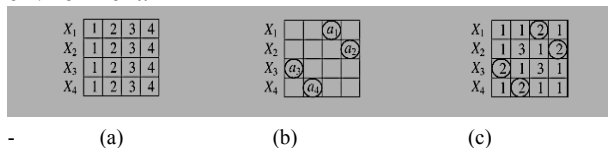


Fig. 6. (a) The representation of domain values for a 4-queen problem. (b) Four agents dispatched into the 4-queen environment. (c) Updated violation numbers corresponding to the positions of the four agents.



Fig. 7. (a) Violation numbers at the initialization step. (b) Violation numbers updated having placed a1 at (3, 1).

At the initialization step, the domain values will be recorded as  $e(i, j).value$  (figure. 6(a)) and the violation numbers for all positions will be set to *zero* (figure 7(a)). After that, agents will be randomly placed into different rows. For instance, if agent  $a_1$  is placed at position (3, 1), the violation numbers in the environment will be updated accordingly, as shown in figure 6(b).

ERA was tested in several applications like n-queen problems and coloring problems and compared with earlier algorithms. Although successful, ERA suffers from a lack of explicit communication and cooperation mechanisms.

## 6. Conclusion

We presented an open community SMA architecture that allows receiving several types of agent societies. The objective is to develop resolution models for constraint satisfaction problems and optimization. Two communities were studied. The first, not described here, relates to the implementation of a deliberation process. The second implements a resolution using the coalition.

The coalition approach permits all agents to participate and to treat proposals, which guarantees to select admissible solution according to the Pareto's optimum. The negotiation provides an environment to obtain solutions by coalition and to avoid the problem's complexity.

An implementation of our approach based on BDI agents made possible to check certain assumptions and to adapt certain resolutions. Agents BDI constitute a favorable environment to express constraints and objectives widely. Future work will relate to the development of the supervisor's role in the interpretation and the backtracking of certain solutions. We will lean towards the use of other interaction modes between the agents for resolution such as the concept of emergence or argumentation.

## References

- [1] Sechen C. Chip planning, total placement, and routing of macro/custom concealment integrated circuits using simulated annealing. In 25th Design Automation Conference, 1988, pp. 73-80.
- [2] Petit T, Poder E. Global propagation of side constraints for solving over-constrained problems, Annals of Operations Research. 184(1), 2011, pp. 295-314.
- [3] Sadgal M. Contribution to the problems of placement and routing. PhD Thesis, Claude-Bernard, University, Lyon 1, 1989.

- [4] Sapena O, Onaindia E, Garrido A, Arangu M. A distributed CSP approach for collaborative planning systems. Original Research Article, Engineering Applications of Artificial Intelligence, 21(5), 2008, pp. 698-709
- [5] Xidonas P, Ergazakis E, Ergazakis K, Metaxiotis K, Askounis D, Mavrotas G, Psarras J. On the selection of equity securities: An expert systems methodology and an application on the Athens Stock Exchange. Expert Systems with Applications. 36 (9), 2009, pp. 11966-11980.
- [6] Hamadi Y, Jabbour S, Sais L. ManySAT a parallel SAT solver. In Journal on Satisfiability, Boolean Modeling and Computation, JSAT, IOS Press, 6(Spec. Issue on Parallel SAT), 2009, pp. 245-262.
- [7] Hsairi L, Ghedira k, Alimi AM, BenAbdelhafid A. Argumentation Based Negotiation Framework for MAIS-E2 model. Chapter VI in book : Open Information Management: Applications of Interconnectivity and Collaboration, Tampere Univ, ISBN:978-1-60566-246, 2009.
- [8] Baykan C, Fox M. Constraint satisfaction techniques for space planning. In Intelligent CAD Systems III - Practical Experiment and Evaluation, 1991, pp. 187-204.
- [9] Golomb S, Baumert L. Backtrack programming. J. ACM, 12, 1965, pp. 516 - 524.
- [10] Shahookar K., Mazunder P. Technical VLSI concealment placement. ACM Computing Surveys, 23 (2), 1991.
- [11] Yokoo M. Algorithms for distributed constraint satisfaction problems: With review. Autonomous Agents & Sys Multi-Agent. 3, 2000, pp. 198-212.
- [12] Mailler R, Lesser VR. Asynchronous Partial Overlay: A New Algorithm for Solving Distributed Constraint Satisfaction Problems. 25, 2006, pp. 529-576
- [13] Modi PJ, Shen W, Tambe M, Yokoo M. An Asynchronous Supplement Method for Distributed Constraint Optimization, Proc. Autonomous Agents and Systems Multi-Agent, Melbourne, Australia, 2003, pp. 161-168.
- [14] Georé JP, Edmonds B, Glize P. Making Coil-Organizing Adaptive Multi-Agent Systems Work - Towards the engineering off emergent multi-agent systems, Methodologies and Software Engineering for Systems Agent, F. Bergenti, M-P. Gleizes, and F. Zambonelli, editors, Kluwer Publishing, 2004.
- [15] Guerra-Hernández A, El Fallah-Seghrouchni A, Soldano H. Distributed Learning in Intentional BDI Systems Multi-Agent, in Proc. ENC, 2004, pp. 225-232.
- [16] Pour HD, Nostary M. Solving the facility and layout and hiring problem by ant-colony optimization-meta heuristic. International Newspaper of Research Production. 44(23), 2004, pp. 5187-5196.
- [17] Boussaa J, Sadgal M. A cognitive Agent for solving problems of occupation of space. Proceedings of the 3rd International Conference on Communications and information technology, December 29-31, 2009, Vouliagmeni, Athens, Greece, pp. 146-152.
- [18] Lawson B, Park S. Asynchronous Time Evolution in year Artificial Society. Newspaper of Artificial Societies and Social Simulation. 3(1), 2000.
- [19] Roy B, Bouyssou D. Multicriterion Assistance with the Decision: Methods and Case. Edition Economica. 1993.
- [20] Jennings NR, Faratin P, Lomuscio AR, Parsons S, Wooldridge M, Sierra C. Automated negotiation: Prospective customers methods and challenges. Intl Newspaper of Group Decision and Negotiation, 10 (2), 2001, pp. 199-215.
- [21] Rahwan I, Ramchurn SD, Jennings NR, Macburney P, Parsons S, Sonenberg L. Argumentation-based negotiation. 18 (4), 2003, pp. 343-375.
- [22] Rubinstein A. Perfect equilibrium in The Knowledge Review Engineering, has bargaining model. Econometrica. 50, 1982, pp. 97-109.
- [23] Faratin P, Sierra C, Jennings NR. Using similarity criteria to make trade-offs in automated negotiations. Artificial Intelligence, 142 (2), 2002, pp. 205-237.
- [24] Amgoud L, Dimopoulos Y, et Moraitis P. With unified and general framework for argumentation-based negotiation. Proc. 6th Intl. Joint Conf on Autonomous Agents and Multi-Agent Systems (AAMAS' 07), 2007, Hawaii.
- [25] Caillou P, Aknine, S, Pinson S. How to Form and Restructure Multi-agent Coalitions. National Conference on Artificial Intelligence (AAAI 02) Workshop on Coalition Formation, Edmonton, Canada, AAAI Press, 2002, pp. 32-37.
- [26] Picard G, Glize P. Model and Analysis of Local Decision Based on Cooperative Self-Organization for Problem Solving. Multiagent and Grid Systems (MAGS), 2(3), 2006, pp. 253-265.
- [27] Liu J, Jing H, Tang YY. Multi-agent Oriented Constraint Satisfaction. Artificial Intelligence, 136(1), 2002, pp. 101-144

**Jamila Boussaa** : Is a Ph.D student in Computational Intelligence and Constraint Satisfaction Problems at Cadi Ayyad University. She received her B.A and DESS degree in computer science from Cadi Ayyad University in 2006. From 2007 to 2009, she was a a web technologie engeneer for SQL Group, and she is currently a trainer for banking system for HPS Solution

**Mohammed Sadgal** : Received the Ph.D degree in computer Science from the university of Lyon in 1989. He received the Ph.D degree in computer vision in 2005. He is currently a Professor (Since 2002) at Cadi Ayyad University (Marrakesh, Morocco). From 1985 to 1987 he was Leader engineer for Net, CAD and CAM from 1987 to 1994 at CONCEPT Society (France). His research interests include Computer Vision, Artificial Intelligence and Multi-agent Systems.

**Aziz Elfazziki** : Received the Ph.D degree in computer Science from the university of Nancy in 1985. He received the Ph.D degree in Multi-agent Systems from Cadi Ayyad University. in 2002, His research interests include Information Systems and Multi-agent Systems.

# Tradeoff Analysis of Bit-Error-Rate (BER) in Cognitive Radio Based on Genetic Algorithm

Shrikrishan Yadav<sup>1</sup>, Prof. Krishna Chandra Roy<sup>2</sup>

<sup>1</sup> Computer Science and Engineering Department, PAHER University  
Udaipur, Rajasthan, India

<sup>2</sup> Electronics and Communication Engineering Department, PAHER University  
Udaipur, Rajasthan, India

## Abstract

We know that the radio electromagnetic spectrum is a natural resource and efficient usage of the inadequate natural resource is one of the greatest challenges of today's wireless communication system just like petrol, coal and water. The efficient use of the available licensed spectrum is becoming more and more critical with increasing demand and usage of spectrum, so this is an urgent need and requirement for the rapidly increasing number of wireless users and also for the conversion of voice oriented applications to multimedia applications. With efficient spectrum use, there are some parameters which also play an important role in the efficiency and performance of a system. One of them is Bit Error Rate (BER), the analysis i.e. optimization (minimization) of BER get better results. The aim of this study is to analysis BER theoretically as well as practically and shows the comparisons between them. In this paper, we analysis BER for better efficiency, high performance and maximum throughput in cognitive radio system based on Genetic Algorithm (GA).

**Keywords:** *Radio spectrum, Cognitive Radio, Bit Error Rate (BER), Optimization, Genetic Algorithm (GA).*

## 1. Introduction

Just like petroleum, wood, water and coal, the natural frequency spectrum is limited and needs to be use more judiciously in order conserve it. It is clear that current static frequency allocation schemes cannot accommodate demands of the rapidly increasing number of higher data rate devices. Therefore; dynamic usage of the spectrum must be distinguished from the static usage to increase the availability of frequency spectrum. For this purpose, Cognitive Radio is proposed as a new technology that provides optimum satisfaction of user requirements like effective spectrum usage and also the smart and secure communication environment.

The transmission scheme of primary users not only occupy licensed bands in frequency, time and space, but also creates a problem with secondary user in a more complicated and structured manner. This problem is removed by cognitive radio which is aware of its environmental, internal state, and location. The radio autonomously adjusts its operations to achieve designed objectives (Mitola 2000). The another way of explaining, the cognitive radio is that it first senses its spectral environment over a wide frequency band, and then adapts the parameters to maximize spectrum efficiency with high performance, while co-existing with legacy wireless networks (Haykin 2005). There are some parameters which affect the performance of the communication system are like power or energy, bit error rate, data rate, bandwidth and channel capacity etc.

## 2. Cognitive Radio

The term, cognitive radio, can formally be defined as follows (FCC Report 2002):

"Cognitive Radio is a radio for wireless communications in which either a network or a wireless node changes its transmission or reception parameters based on the interaction with the environment to communicate efficiently without interfering with licensed users."

The cognitive capability of a cognitive radio enables real time interaction with its environment. This interaction helps to determine the appropriate communication parameters in order to adapt the dynamic radio environment. The radio analyzes the spectrum characteristics and changes the parameters at real time to

provide a fair scheduling among the users that share the available spectrum. With this approach to solve the issue of scarcity of available radio spectrum, the Cognitive radio technology is getting a significant attention. The primary feature of cognitive radio is the capability to optimize the relevant communication parameters given at a dynamic wireless channel environment.

There have been implementations of GA based cognitive radio implementations, but the performance of these algorithms has not been thoroughly analyzed. The fitness functions employed in these algorithms have also not been explored in detail. Specifically, the analysis comes from finding the non-dominated solutions in the solution space, which is known as Pareto front. Genetic algorithms (GA) are used to optimize multi-objective problems, and can produce the Pareto Front. After the Pareto front has been optimized, the final challenge is to make a decision about the waveform on the Pareto front which stands for Quality of Service satisfaction.

### 3. Bit Error Rate (BER)

Bit error rate (BER) of a communication system is defined as the ratio of number of error bits and total number of bits transmitted during a specific period. In digital transmission or digital communication system, the number of bit errors is the number of received bits of a data stream over communication channels that have been altered due to noise, interference, distortion or bit synchronization errors in the system. The bit error rate or bit error ratio (BER) is the number of bit errors divided by the total number of transferred bits during a considered time interval. BER is a unit less performance measure, often expressed as a percentage (%).

As an example, let assume 10 bits of data is transmitted as a bit sequence:

0 1 1 0 0 0 1 0 1 1,

and suppose the following bit sequence is received at receiver side:

0 0 1 0 1 0 1 0 0 1,

The number of bit errors (the underlined bits) in this case is 3. The BER is 3 incorrect bits divided by 10 transferred bits, resulting in a BER of 0.3 or 30%.

a) Factors affecting the BER

In a communication system, the receiver side BER may be affected by:

- Transmission channel noise.
  - Interference.
  - Distortion.
  - Bit synchronization problems.
  - Attenuation.
  - Wireless multipath fading, etc.
- b) The BER may be improved by:
- Choosing strong signal strength (unless this causes cross-talk and more bit errors).
  - Choosing a slow and robust modulation scheme or line coding scheme.
  - Applying channel coding schemes such as redundant forward error correction codes.

It is the likely that a single error bit will occur within received bits of independent rate of transmission. There are many ways of reducing BER. Here, we focus on channel coding techniques.

A channel in mobile communications can be simulated in many different ways. The main considerations includes the effect of multipath scattering, fading and Doppler shift that arises from the relative motions between the transmitter and the receiver. In our simulations, we have considered the two most commonly used channels: the Additive White Gaussian Noise (AWGN) channel where the noise gets spread over the whole spectrum of frequencies and the Rayleigh fading channel.

### 4. Genetic Algorithm

Genetic algorithm (GA) is the technique based on evolutionary computation to find approximate solutions to the optimization problems. Genetic algorithms are inspired by the Darwin's theory of evolution which is best or simply the survivor among the available pool is an evolved solution. The evolutionary computation may involve techniques like inheritance, mutation, selection and crossover to provide the best possible optimization.

In 1992 John Koza introduced "Genetic Programming" (G.P.). Since the introduction, the G.A's have been used to solve difficult problems like, Non deterministic problems and machine learning as well as the evolution of simple programs, pictures and music. The main advantage of Genetic Algorithm over the other methods is their parallelism. It travels and search spaces that use more individuals for the decision-making so they are less likely to get fixed in a local extreme like other available decision-making techniques. The GA uses a population of

chromosomes that represent the search space and determine their fitness by a certain criterion (fitness function). In each generation (iteration of the algorithm), the most fit parents are chosen to create offspring, which are created by crossing over portions of the parent chromosomes and then possibly adding mutation to the offspring.

The Genetic algorithms approach is used for the optimization of the decision- making module in the radio. They are well suited to the multi-objective functions due to their convergence behavior towards the optimized solution and help the radios in adaptation for the decision-making process. Apart from this, the genetic algorithms also provide the optimization in decision making with multiple advantages. They provide flexibility in problem analysis, as long as the chromosome and the objective functions are defined properly. The convergence behavior of the genetic algorithm is really helpful in our application, i.e. the Cognitive Radios. This algorithms may have a long convergence time for an optimal solution but normally do not take much time to give very good solutions [7].

Outline for the Genetic Algorithms:

1. Start: Generate a random initial population of  $n$  chromosomes that consists the available solutions for the problem.
2. Fitness: Imitate the fitness of each of the chromosomes in the initial population.
3. New population: Reproduce, according to the following steps until the next generation completes.
4. Selection: Select two chromosomes that have the best fitness level among the current population.
5. Crossover: In this step two selected chromosomes considering the crossover probability are crossover, to form the off springs for the next generation. If this operation is not performed the children would be the exact copy of the parent chromosomes.
6. Mutation: Transform the new offspring at each defined mutation point, considering the mutation probability and places it in the new population.
7. End Condition: Repeat the above steps until certain condition (maximum no of population or the desired optimum has been reached), has been met. [7]

A representation for the chromosome must provide the information about the solution that it represents. The most popular of all representations is the binary string. Where

each bit in the string can represent the chromosome characteristics or the whole string cumulatively can do this. The use of integer or real number representations for the chromosomes can also be useful.

## 5. Simulations and Results

In the real world problem such as the problem arrangement with this paper, the found solutions solve the objective solutions even when they are conflicting, that is, minimizing one function may also degrade other functions. For example, minimizing BER and minimizing power simultaneously generate a divergence because of the single parameter i.e. transmit power, which affects each objective in a different manner. Obtaining the optimal set of decision variables for a single objective minimizes power and often the outcome is non-optimal set with respect to other objectives, e.g. minimize BER.

In this work, the results on BER versus  $E_b/N_0$  are obtained both theoretical and genetic algorithms. Using simple CR parameter as shown in Table 2 can be obtained as Bit-Error-Rate (BER) and corresponding Signal-to-Noise Ratio (SNR). The length of chromosome utilized by proposed GA has been shown in Fig. 1 which represents individual in the population. The genetic parameter used for GA shown in Table 3. The simulation results were observed by several hundred times.

Table 1: Values of Cognitive Radio Environmental Parameters

Parameter	Symbol	Min. Value	Max. Value	Step Size
Noise Power	N	-110dBm	-25dBm	-1dBm

Table 2: Values of Cognitive Radio Transmission Parameters

Parameter	Symbol	Min. Value	Max. Value	Step Size
Transmit Power	P	-35dBm	35dBm	1dBm
Bits in each Symbol	K	2	2	
Bandwidth	B	1Mhz	10Mhz	1Mhz
Symbol Rate	R	1Mbps	8Mbps	1Mbps



Table 3: Genetic Parameter Settings

Genetic Parameters	Proposed
Population Size	20
Maximum generation	100
Crossover type	Single-point
Crossover rate	0.8-0.95
Mutation type	Randomly change the integer value within given range
Mutation rate	0.1

P	K	B	L	R	N
---	---	---	---	---	---

Fig. 1 Chromosome Lengths

Genetic Algorithm has been implemented using Eq. (1). The theoretical and GA computational values have been recorded as shown in Table 4. The coordinate axes are used to represent the comparison between BER and SNR. The y-axis represents the score for BER, while the x-axis is the score for the ratio of the energy per bit ( $E_b$ ) to the noise power spectral density ( $N_0$ ).

$$P_{be} = \frac{2}{k} Q \left( \sqrt{2k\gamma} \sin \frac{\pi}{M} \right) \quad (1)$$

The parameter  $x$  corresponds to the decision vector of variable used as inputs to the fitness functions. For every curve, as the fitness score for BER objective decreases, the value for the  $E_b/N_0$  increases. This trade-off analysis has to be made by using optimization function. In this paper, fitness functions using the defined set of parameters has been developed, that are used by Cognitive Radio engines to establish a single optimal transmission parameter result to get the optimal solution.

Table 4: Theoretical and GA Comparison of BER

$E_b / N_0$ (dB)	Theoretical BER	GA BER
-15	0.400	0.380
-10	0.330	0.320
-5	0.213	0.213
0	0.079	0.078
5	0.006	0.059
10	$3.87 \times 10^{-6}$	0.000

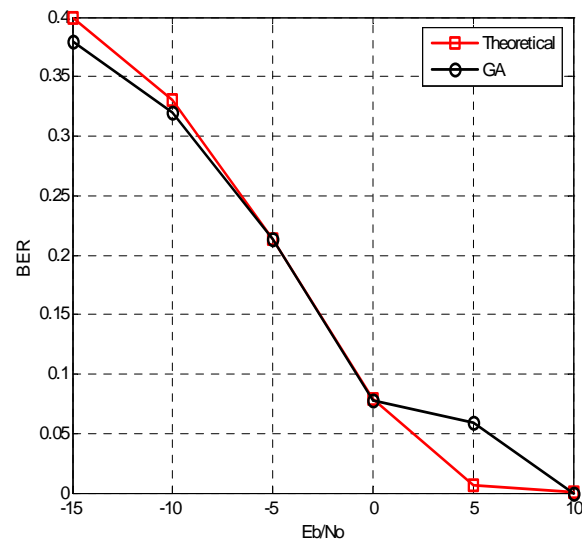


Fig. 2 Theoretical and GA Comparison of BER

## 6. Conclusion

In this paper the contributions has been made in the area of Cognitive Radio. The research work achievements of this hypothesis are following: Multi-objective cost functions, which set up the relationships between the environmental parameters, transmission parameters, and objectives of Quality of Service performance, were developed. This work discovered that the Genetic Algorithm based system approach was more robust and offers an interface that permits the user to easily adjust

Cognitive Radio parameters such as BER which increase the performance, reliability and capability of the system.

## References

- [1] D. Cabric, S. M. Mishra, and R. Brodersen, "Implementation issues in spectrum sensing for cognitive radios," in Proc. 38th Asilomar Conf. Signals, Systems and Computers, Pacific Grove, CA, Nov. 2004, pp. 772–776.
- [2] I. F. Akyildiz, W. Y. Lee, M. C. Vuran, S. Mohanty, "NeXt generation dynamic spectrum access cognitive radio wireless networks: A survey," *Computer Networks*, 50, 2006, pp 2127-2159
- [3] FCC, "Spectrum policy task force report," ET Docket No. 02-155, Nov. 2002.
- [4] Joint Tactical radio Systems, "Software communications architecture specification," November 2002.
- [5] R. Etkin, A. Parekh, and D.Tse, "Spectrum sharing for unlicensed bands," in IEEE International symposium on New Frontiers in Dynamic Spectrum Access, 2005, pp 251-258
- [6] Spectrum Policy Task Force, "Report of the spectrum policy workgroup," November 2002. [Online]. Available: [http://www.fcc.gov/sptf/files/SEWGFfinalReport\\_1.pdf](http://www.fcc.gov/sptf/files/SEWGFfinalReport_1.pdf)
- [7] C.J. Rieser, "Biologically inspired cognitive radio engine model utilizing distributed genetic algorithms for secure and robust wireless communications.
- [8] Digham,F., M. Alouini, and M. Simon. 2007. On the Energy Detection of Unknown Signals Over Fading Channels *IEEE Transactions on Communications* 55: 21-24
- [9] Federal Communications Commission's, "Spectrum policy task force report (ET Docket No. 02-135)," Nov. 2002.
- [10] S. Haykin, "Cognitive radio: brain-empowered wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 23, pp. 201–220, Feb. 2005.

**Shrikrishan Yadav:** working as an Assistant Professor in Computer Science and Engineering Department in PAHER University, Udaipur, India. He has completed B. E. in computer science and engineering from Mohanlal Shukhadia University, Udaipur and pursued M.Tech. in information communication from Gyan Vihar University, Jaipur. He has more than two years of teaching experience. He is also presented and published 9 papers in National and International journals and conferences. He is an associate member of Computer Society of India (CSI) and a member of International Association of Engineers (IAENG). His current research interest includes Cognitive Radio, Wireless Sensor Networks, Artificial Intelligence, Information Communication etc.

**Prof. Krishna Chandra Roy:** received his M.Sc. (Engg.) degree in from NIT Patna, Bihar, India and Ph.D degree in "Digital Signal Processing in a New Binary System" year 2003. He has currently professor & Principal in Pacific College of Technology Udaipur, India and having more than 15 year teaching and research experiences. He guided many Ph.D scholars. He is also published and presented more than 55 papers in International and National journals and conferences. He published two books Problems & solution in Electromagnetic Field Theory by Neelkanth Publishers (p) Ltd., Year-2006 and Digital Communication by University Science Press, Year-2009 respectively. His current research interests include Digital Signal Processing and wireless embedded system.

# Evaluation Strategy for Ranking and Rating of Knowledge Sharing Portal Usability

D.Venkata Subramanian, Angelina Geetha

Department Of Computer Science and Engineering  
B.S. Abdur Rahman University, Chennai, India

## Abstract

Knowledge Sharing portals are the primary gateways for users to access all the information they need for their activity with adequate safety, security and in the best quality. The usability of a Knowledge Sharing Portal plays an important role in every organization and higher learning institutions, as it helps to increase the user satisfaction, reuse of knowledge assets, consistency of information and simplification of the maintenance process, regardless of the context, order or type of users. The primary goal of this paper is to propose a strategy for the ranking and rating of usability of the KM system. In this proposed work, we first describe the ways to design and develop the quality factors, using a multi-dimensional metric model for measuring usability along with other supporting factors. Secondly we have shown the ways to apply the Weighted Average Mean (WAM) on the usability factor and other relevant factors for evaluation. Using the weights and values generated in the metric database, the usability of the KM system can be ranked and rated both manually and automatically.

**Keywords:** *Multi-dimensional Metric Model, Knowledge Sharing Portal, Ranking, Rating, Usability Evaluation, WAM*

## 1. INTRODUCTION

Knowledge Management (KM) provides an innovative methodology for knowledge creation, storage, dissemination and sharing. Many companies and institutions are utilizing KM systems, especially the Knowledge Sharing Portal as the main method of collaboration for increasing, knowledge sharing with their workers. The usability evaluation of the KM portal is one of the crucial steps, if an organization wants to change the structure of the portal. The evaluation methodology also helps to identify the frequently used as well as effective documents with higher usability and at the same time, identify least recently used contents or inactive contents for archiving them for better storage utilization. The feedback from users helps to identify areas where access mechanisms, structure, labeling and depth of content need to be improved to meet the user's needs. Considering the intangible nature of the knowledge asset, complexity and dynamics of building the KMS infrastructure, one of the possible approaches is to determine the strengths and weaknesses of the existing KM portal and its components, such as contents through usability evaluation. Speed of information change and new ways of collaboration and enhanced user interfaces have started to take place, at

every organization for knowledge sharing and management. Willingness and accurate inputs from the knowledge seekers or providers will decrease, if the feedback is requested many times from the system or through a manual process. So it is important to enhance or develop a suitable strategy for usability evaluation of the Knowledge Sharing Portal to reduce the number of user feedbacks and at the same time extract the maximum results from a metric database.

Based on many research works, it has been identified that there is no proven reliable model and metric database, to estimate and report the usability of the Knowledge portal. To overcome this challenge, we have developed the multi-dimensional metric model [3] and the widely used statistical technique WAM to rank and rate the system-generated and user measures, which are stored in the metric database. This paper first describes the comprehensive Knowledge Management Systems framework [2], usability of the KM portal, and the prediction process of using multiple dimensions. Secondly, the paper describes the process of evaluating usability through experiments and approaches for building a metric database using the multi-dimensional metric model for capturing the measures and metrics. The WAM will be applied against the results to rank and rate the effectiveness of the KM portal.

## 2. KM PORTAL AND USABILITY

A Knowledge Management (KM) system is a collective term that is used to describe the creation of knowledge repositories, with their respective interface components, improvement of knowledge access and sharing as well as communication through collaboration, enhancing the knowledge environment and managing knowledge as an asset for an organization. Considering the fundamental capabilities of the KMS and typical KMS infrastructure topology, we have identified a suitable KMS framework [2] which is mentioned in the figure (Figure 1) below. This framework represents all the components which make up the KMS, and in particular focuses on the needed quality factors. For our research work, we have taken Usability as an important quality factor, and other supporting factors such as Availability, Functionality and

Efficiency, for evaluating the effectiveness of the Knowledge Sharing Portal, in totality.

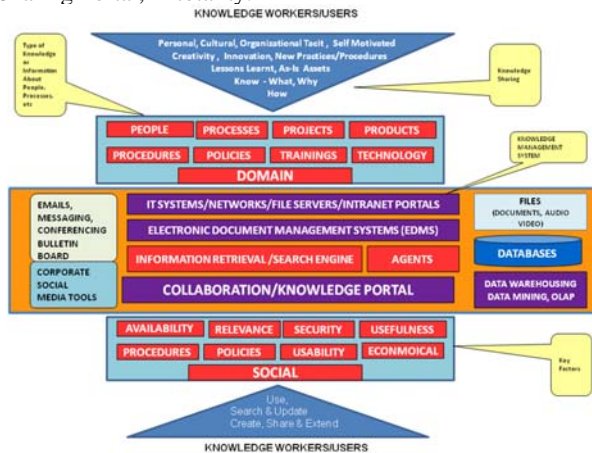


Figure 1. KMS Framework

The Knowledge Portal works as an integration tool to provide easy, unified and integrated access to an organization's own resources. Most knowledge portals have existing, but diverse systems for collecting and accessing important information from all the different systems or groups. An effective knowledge portal would provide a single point of access to all of the systems and would be structured in such a way that the location and retrieval of such information would be quick and easy. Knowledge Portal helps as an access tool for other information sources to provide internal and external information, which are beyond their own organization's resources and which can be made available to staff. The Knowledge Portal also serves as a communication tool to enable individuals, teams and communities of practice to share and discuss ideas and knowledge.

The International Organization for Standardization (ISO) defines the Usability of a product as "the extent to which the product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use."

For our evaluation, we consider the following key parameters for evaluating the usability of a KM System

- Learnability, User Efficacy, User Efficiency
- Searchability, Memorability, Operability
- Communicativeness, Accessibility
- User satisfaction and Expert Ranking

### 3. EVALUATION PROCESS

Based on our earlier research work, we had proposed a hybrid method of using the Goal Question Metric (GQM) and Balanced Score Card to collect the required measures for performing an evaluation [2]. As illustrated in Figure 2, the basic KM System prediction process consist of the selection of

the quality dimensions and their classification in to subjective or objective and then applying the hybrid method for data collection. The selected measures which are generated manually or through the system are stored in any available database and later retrieved for ranking and rating using the WAM Method. The factors whose effects need to be quantified are called primary quality factors. The features often discussed concerning the overall quality of the knowledge management system are capability, availability, reliability, usability, maintainability and completeness. In our metric model design and development, these quality factors are considered as dimensions.

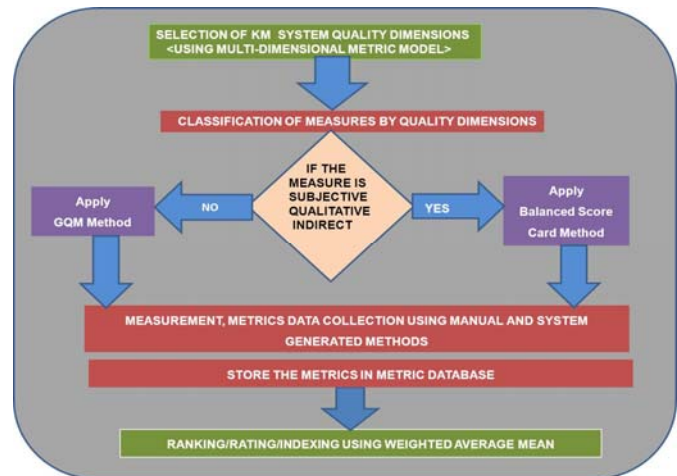


Figure 2: Hybrid Evaluation Approach

We categorized the factors in two groups such as primary and secondary and evaluators can choose the factors based on the evaluation consideration and outcomes. Continuous Inputs for the KMS Measurement process can be done through standard system programs or tools to monitor the usefulness and responsiveness of supporting technology or the framework or components used for the KMS. They give an indirect indication of knowledge sharing and reuse by highlighting which knowledge assets are the most popular ones and which components are mostly accessed and used by the knowledge workers. The system generated factors may also indicate usability problems and supporting policies for the KMS by introducing the agents which collect these measures. Some of the standard ones are page visits, number of community members and size of the document or file system.

The output metrics can be calculated or derived based on the objective and subjective feedback analysis from the knowledge sharing portal or the whole knowledge infrastructure. Most of these output measures can be calculated manually or through online user survey or forced feedback system of the usage of the knowledge portal. Some of these measures can be calculated using system level statistics and also by developing some background agents or web services. For our evaluation we have considered the most popular usability review checklist supplied by Xerox Corporation and usability evaluation questionnaires can be decided based on the usability requirements set by the evaluator of the KM portal.

#### 4. METRIC DATABASE

The metric database Entity Relationship (ER) diagram is shown in the figure (Figure 4) below, it was created to hold the user and expert feedback of the considered dimensions(4) and measures for evaluating the KM portal usability. The databases used in the existing infrastructure can be considered for storing the metrics and measurements. As the volume of data and the amount of transactions used for the KMS measurement are less, there is no need for a dedicated or high performance database and the existing database used for infrastructure maintenance or application database can be used to store the schema and data.

The metric database can be created using any industry specific database systems, using the following steps:

1. Gather the evaluation factors for assessment.
2. Decide the Quality Factor and Sub Factors
3. Create Data Objects such as Tables or Classes or XML to hold the Quality Factor/Entities and Attributes/measures.
4. Upon collecting the measures, store them in the data objects

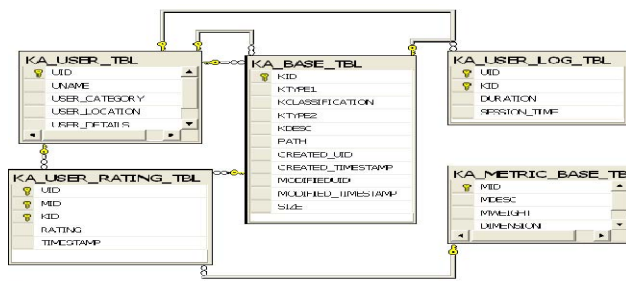


Figure 4: Metric Database E-R Diagram

As you see from the above Entity Relationship E-R diagram, our metric database has been designed in such a way as to be flexible to hold any dimension and metric as per the rating from the user (normal user or expert user or from the system).

The database consists of five key tables, namely

- The KA\_BASE\_TBL which is the base table which contains the knowledge asset created/modified in the knowledge portal or repository.
- The KA\_USER\_TBL holds the information about the user name, user type and their details.
- The KA\_USER\_RATING table holds the user feedback on the given measure.
- The KA\_METRIC\_BASE\_TBL which is our key table that holds attributes, like the metric id (MID), metric Description (MDESC), metric weight (MWEIGHT) and quality DIMENSION.

#### 5. DIMENSIONS AND MEASURES

The following section discusses the measures and metrics corresponding to some of the prime quality factors, which will

be given weights in the 80% category. The diagram below represents the four key dimensions considered for the evaluation of the KM systems. The needed dimensions and attributes can be added as per the evaluation or prediction. For evaluating the usability of the knowledge management system, one needs to consider Usability as the primary dimension and the other dimensions such as Efficiency, Availability, Functionality, etc as secondary.

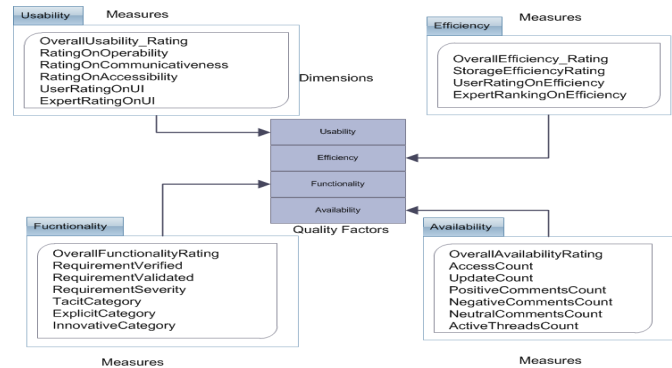


Figure 5: Four dimensional evaluation model

##### 5.1 Usability Measures

The knowledge component should be easily understandable, learnable, and applicable. Usability attributes are the features and characteristics of the software/product/sites that influence the effectiveness, efficiency and satisfaction with which users can achieve specified goals. The usability entity can have the following attributes:

- OverallUsability\_Rating
- RatingOnLearnability
- RatingOnUserEfficacy
- RatingOnUserEfficiency
- RatingOnSearchability
- RatingOnMemorability
- RatingOnOperability
- RatingOnCommunicativeness
- RatingOnAccessibility
- UserSatisfactionRatingOnUI
- ExpertSatisfactionRatingOnUI

##### 5.2 Functionality Measures

The functionality of the KMS can be considered as an entity object in the metric model and its expected behavior will be captured as the requirements or attributes of an entity object

- OverallFunctionalityRating
- RequirementVerified
- RequirementValidated
- RequirementSeverity
- TacitCategory
- ExplicitCategory
- InnovativeCategory

### 5.3 Availability Measures

In the context of the KMS, Knowledge Availability is whether (or how often) a given knowledge asset is available for use by its intended users. The following are some of the key attributes for measuring the Availability of the knowledge asset:

- OverallFunctionalityRating
- RequirementVerified
- RequirementValidated
- RequirementSeverity
- TacitCategory
- ExplicitCategory
- InnovativeCategory

### 5.4 Efficiency Measures

The knowledge portal should state the quickest solution with the least resource requirements. The following describes the efficiency of the KMS components and their contents:

- OverallEfficiencyRating
- StorageEfficiencyRating
- UserRatingOnEfficiency
- ExpertRankingOnEfficiency

## 6. WEIGHTED MEAN

The weighted mean is similar to an arithmetic mean (the most common type of average), where instead of each of the data points contributing equally to the final average, some data points contribute more than others. The weights can be specified for the measures or metrics collected, based on the user or implementer or evaluation needs. For example, training organizations gives higher weightage to the quality of the training material, certifications obtained by the participants and also the employability of the participants after completing the training which indicated its effectiveness. Consulting companies usually gives weights on the problems solved by the consultant or technical expert and the number of best practices implemented by the technical or domain consultant on a given domain or technical issue. In higher learning institutions like universities and engineering colleges, the relevance of the knowledge asset based on the syllabus or curriculum gets more weight as it reflects the standard. Additional weights can be given to the academic institutions, if the knowledge asset on a specific subject encourages students to do projects or submit research papers or obtain higher scores in the exams in the subjects taught and presented in the knowledge portal or repository. The non-critical or supportive measures, which indirectly contribute to measurement of the KM System, will be considered as non-weighted measures for our evaluation. 95:5 has been adapted for our ranking and rating using the weighted average mean, in these formulas, 90 indicates 95% of the allocation for critical KMS measures and 5 indicates 5% of the allocation for non-weighted measures. We have considered the following as the non-weighted measures, and the evaluator can decide the rule

and keep the measures in the weighted or non-weighted category.

- KM Infrastructure & Management Support
- Participant's Subject or Domain Knowledge
- Participants' Thrust for Knowledge Collaboration

## 7. USERS AND ENVIRONMENT

For evaluating the usability of the KM portal documents, the metric must indicate the usability parameters of the KM portal and/or its components and need to involve Knowledge Seekers, Providers, Web Portal Designer, Developer and Domain expert on the evaluation subject and typical novice and expert web users. The participants were selected from multiple departments of top two engineering colleges with moderate and frequent usage of knowledge portals and knowledge repositories. Initially 450 candidates were selected, but only 270 candidates, with similar profiles, were actually used in the experiment.

The main aspects of the normal user profiles of the participants were similar in the following ways:

- Computing knowledge and knowledge of using collaborative tools such as corporate or organization knowledge portal or knowledge systems
- $\geq 22$  years of age and  $< 27$  years of age, with English as their learning language for all the engineering subjects.

The main aspects of the expert user profiles of the participants used were similar in the following ways:

- Teaching or Training skills
- Expert knowledge in the subject area
- Willingness to review and provide ranking of the knowledge asset in a constructive way
- $\leq 69$  years of age and  $> 27$  years of age with apt qualification and teaching experience in the engineering subjects

The users evaluated six knowledge sharing portals for the purpose of usability evaluation. One of the categories is technology forum/support site for IT tools and database related support; another category is engineering colleges and the third category is training firms. The data obtained for this experiment pertained to the usability of the computer science related subjects and respective user interfaces provided in the knowledge sharing portal.

## 8. RANKING AND RATING

Though we have considered Usability as a primary dimension, we have also considered other dimensions as important which should be evaluated in conjunction with the usability of the Knowledge Sharing Portal. The knowledge assets and the portal should be available and part of functionality and efficiency requirement of the KM Infrastructure. The table 1 shows how the evaluation percentage is distributed among multiple dimensions.

TABLE 1. DISTRIBUTION TABLE FOR EVALUATION

KM Quality Dimension For Evaluation	Allocated Evaluation Percentage
Usability	80%
Functionality	5%
Availability	5%
Efficiency	5%
Others	5%
	100%

The following table 2 shows how the evaluation percentage (80%) is distributed among multiple usability measures and the captured rating received from the metric database and weighted calculation for the usability dimension.

TABLE 2. WEIGHTAGE TABLE FOR USABILITY

KM Usability Measure	Weight (80%)	Captured Rating (from Metric DB)	Weighted Calculation
RatingOnLearnability	10%	3.5	0.35
RatingOnUserEfficacy	10%	3.9	0.39
RatingOnUserEfficiency	10%	3.7	0.37
RatingOnSearchability	10%	4.4	0.44
RatingOnMemorability	10%	4.1	0.41
RatingOnOperability	10%	3.6	0.36
RatingOnCommunicativeness	10%	4.2	0.42
RatingOnAccessibility	10%	3.8	0.38
UserSatisfactionRatingOnUI	10%	3.7	0.37
ExpertSatisfactionRatingOnUI	10%	4.5	0.45
	<b>100%</b>	<b>Total Score</b>	<b>3.94</b>
		<b>KM Dimension Score</b>	<b>3.152</b>

The secondary tables (table 3 to 6) listed below show how the evaluation percentage (15%) is distributed among other dimensions such as functionality, availability and efficiency.

TABLE 3. WEIGHTAGE TABLE FOR FUNCTIONALITY

KM Functionality Measure	Weight (5%)	Captured Rating (from Metric DB)	Weighted Calculation
OverallFunctionalityRating	20%	3.5	0.7
RequirementVerified	10%	3.2	0.32
RequirementValidated	10%	3.6	0.36
RequirementSeverity	10%	2.1	0.21
TacitCategory	10%	4.1	0.41
ExplicitCategory	10%	2.4	0.24
InnovativeCategory	30%	3.1	0.93
	<b>100%</b>	<b>Total Score</b>	<b>3.17</b>
		<b>KM Dimension Score</b>	<b>0.1585</b>

TABLE 4. WEIGHTAGE TABLE FOR AVAILABILITY

KM Availability Measure	Weight (5%)	Captured Rating (from Metric DB)	Weighted Calculation
OverallAvailablityRating	20%	3.2	0.64
AccessCount	10%	3.9	0.39
UpdateCount	10%	3.2	0.32
PositiveCommentsCount	10%	4.2	0.42
NegativeCommentsCount	10%	1	0.1
NeutralCommentsCount	10%	3.7	0.37
ActiveThreadsCount	30%	4.1	1.23
	<b>100%</b>	<b>Total Score</b>	<b>3.27</b>
		<b>KM Dimension Score</b>	<b>0.1635</b>

TABLE 5. WEIGHTAGE TABLE FOR EFFICIENCY DIMENSION

KM Efficiency Measure/Metric	Weight (5%)	Captured Rating (from Metric DB)	Weighted Calculation
OverallEfficiency_Rating	20%	3.9	0.78
StorageEfficiencyRating	20%	3.5	0.7
UserRatingOnEfficiency	30%	4.1	1.23
ExpertRankingOnEfficiency	30%	3.6	1.08
	<b>100%</b>	<b>Total Score</b>	<b>3.79</b>
		<b>KM Dimension Score</b>	<b>0.1895</b>

The table 7 shows the overall summary weightage for the weighted dimension group and non-weighted dimension group.

TABLE 6. NON-WEIGHTED MEASURES

KM Non Weighted Measure/Metric (5%)	Captured Rating (from Metric DB)
Supporting KM Infrastructure	3
Management Support	2
Conducive Environment	1
Participant's Subject or Domain Knowledge	2
Participants Thrust for Knowledge Collabrator	3
Total Non Weight Value	11
Average	2.2
5 % of Non Weighted Dimension Group Score	0.11

TABLE 7. SUMMARY WEIGHTAGE TABLE

Allocation	Type	Derived Score
90%	Weighted Dimension Group	3.6635
10%	Non Weighted Dimension Group	0.11
	<b>Overall KM System Usability Score</b>	<b>3.7735</b>

TABLE 8. RANKING AND RATING TABLE

Rank	Category	Rating
1	Outstanding	5
2	Extremely Usable	4
3	Usable	3
4	Somewhat Usable	2
5	Not Usable	1

Table 6 indicates the non-functionality measures such as overhead measures for the remaining 5% evaluation. The negative values/rating will be subtracted from the score. The ranking and rating was done based on the data collected from the knowledge portals. The captured ratings are aggregated values, which are stored in the database through system feed, user and expert feed. The ranking and rating values for all the six portals considered are listed below table 9.

TABLE 9: USABILITY RATING OF KM PORTALS

Knowledge Sharing Portal	Usability Score
Portal-1	4.214
Portal-2	2.489
Portal-3	4.561
Portal-4	3.654
Portal-5	3.773
Portal-6	2.542

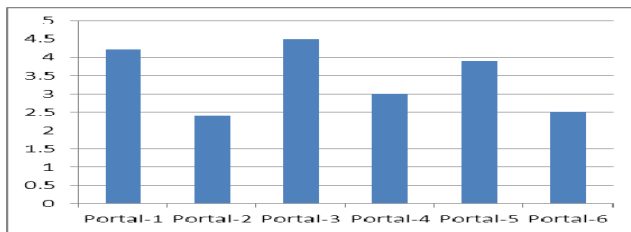


Figure 6: Usability Ranking Of KM Portals

The chart shown in Figure 6, represents the ranking and rating of the KM portals evaluated as part of the experiment.

## 9. CONCLUSIONS

By referring to table 8 and the guideline table 9 for ranking and rating, it is clear that the evaluated KM system usability is effective as the system got the overall evaluation rating of 3.7735. In this research work, we have attempted to use the metric database and a proven statistical technique, the Weighted Average Mean, to validate the effectiveness of the knowledge sharing in Knowledge Management systems using Ranking and Rating and proved that the combination of the Metric Database with a statistical technique such as the WAM could be useful to predict the usefulness and effectiveness of the Knowledge Sharing Portal and can help to identify issues, challenges and gaps in the existing KM infrastructure and could help to improve the user-satisfaction of the Knowledge Sharing Portal. For our experiment, we have taken only a four dimensional metric model and database, to validate the usability which may not be adequate enough to measure the overall usability of the Knowledge Management Systems, as the usability measurement is heterogeneous. So there are multiple research avenues to enhance the proposed dimensional model as well as the structure of the metric databases to use multi-dimensional data cubes instead of tables to gather additional usability factors and measures. The proposed strategy is simple to use and provides a lot of flexibility for evaluators to decide the usability dimensions of a KM system and store them in the database system for reporting, ranking and rating as per the allocation of weights. In order to compare the proposed strategy, an appropriate hypothesis can be set to validate the significance of the results obtained from the KM metric database. Further research can be conducted on mining methods along with ontologies for getting highly reliable, dependent and multifarious relationship of the usability factors and their attributes.

## References

- [1] Dayanandan Venkata Subramanian, Angelina Geetha (2011), "Guidelines and Effectiveness of Using Social Media Tools For Knowledge Sharing Communities", National Conference on Knowledge Intelligence and Telematics, Gujarat, India
- [2] Venkata S Dayanandan, Angelina Geetha(2011), "Adaptation of Goal Question Metric Technique For Evaluation of Knowledge Management Systems", Review of Knowledge Management Journal, Volume 1, Jan-June 2011, ISSN:2231-5845
- [3] Venkata S Dayanandan, Angelina Geetha, Mohammed Hussain(2011), "Measurement Process and Multi-dimensional Model For Evaluating Knowledge Management Systems", International Conference on Research and Innovation in Information Systems, International Islamic University, Malaysia (IIUM) and Universiti Teknologi Malaysia(UTM), Indexed by IEEE & SCOPUS, November 2011
- [4] Diwakar Krishnamurthy, Jerry Rolia, and Min Xu I(2011), "WAM – The Weighted Average Method for Predicting the Performance of Systems with Bursts of Customer Sessions, University of Calgary, Calgary, AB, Canada, IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, 0098-5589/11
- [5] Jiawei Han, Yizhou Sun, Xifeng Yan, Philip S. Yu(2010), "Mining Knowledge from Databases: An Information Network Analysis Approach", SIGMOD'10, June 6–11, 2010, Indianapolis, Indiana, USA.
- [6] Sergio Consoli, Keneth Darby-Dowman, "Heuristic approaches for the quartet method of hierarchical clustering", IEEE Transactions on Knowledge & Data Engineering, Vol 22 #10, 2010
- [7] Behrang ZadJabbari, Pornpit Wongthongtham, Farookh Khadeer Hussain(2010),"Ontology based Approach in Knowledge Sharing Measurement", Communications of the IBIMA Volume 11", 2009
- [8] Estelle de kock, jduy van bilijon, marco pretorius, "Usability Evaluation methods: Mind the gaps", SAICSIT, South Africa, 2009
- [9] Ashleigh-Jane Thompson, Elizabeth A Kemp, "Web 2.0: Extending the framework for heuristic evaluation", Proceedings of the 10th international conference NZ Chaperite of the ACM's special interest group on Human Computer Interface, 2009
- [10] Stephanie D Swindler, Christian E Randall, Gary S Beisner II, "An application of heuristic techniques for evaluation of collaborative systems interfaces", International Symposium on Collaborative Technologies and systems, 2006.
- [11] Rusli Abdullah, Mohd Hasan Selamat, Universiti Putra Malaysia and Shamsul Sahibudin, Rose Alinda Alias(2005), "A Framework For Knowledge Management System Implementation In Collaborative Environment For Higher Learning Institution", Universiti Teknologi Malaysia
- [12] B. Martens and W. Jabi, "Knowledge Portal as a new paradigm for scientific publishing and collaboration", IT conference, 2004
- [13] Nielsen, J. and R. Molich (1990), "Heuristic Evaluation of User Interfaces", New York, NY: ACM Press.

**First Author** D. Venkata Subramanian completed his Bachelor of Engineering in Computer Science and Engineering in 1992 from Bharathidasan University, India and his Master Of Science in Computer Systems Engineering from Northeastern university, Boston, USA in 2002. Venkata has worked in industries in India, Malaysia, Singapore & U.S.A for about 18 years and got certifications in IBM Db2, MDM Infosphere and ITIL V2. Venkata is presently providing consulting and training in the areas of Databases and Systems, and in parallel doing his Ph.D (part-time) in Knowledge Management, at B.S Abdur Rahman University, Tamilnadu, India.

**Second Author** Angelina Geetha completed her Masters In Engineering Degree and Ph.D in Computer Science from Anna University. Angelina is presently working as a Professor and Head in Computer Science and Engineering department at B.S Abdur Rahman University, Chennai, Tamilnadu India. Angelina has worked for more than a decade in academic institutions and guided several Ph.D students in knowledge management, data warehousing, data mining and web mining.



# Vehicular Ad-hoc Networks

**Asad Maqsood, Rehanullah Khan**  
Department of Electrical Engineering  
Sarhad University of Science and IT  
Peshawar, KPK, Pakistan

## Abstract

*Modern day's vehicles require advanced communication system on board to enable passengers benefit the most from available services. IEEE 802.11p is the new extension of IEEE 802.11 standards; especially proposed for the high vehicular environment. The WAVE documentation represents enhancements to the Media Access Control (MAC) and Physical (PHY) layer of IEEE 802.11 standards to work efficiently in high vehicular environment. In this research work, the main emphasis is on the new IEEE 802.11p enhancement of MAC and PHY layers. More specifically, the target of this research is to setup a simulation environment which will allow us to investigate the use of real time voice application, using IEEE 802.11p (WAVE) enhance setting, in a single hop and multi-hop environment where nodes are not directly connected. Also, the evaluation of transmission between moving nodes are tested by simply sending and receiving FTP file between them with varying speed of the moving nodes.*

**Keywords:** VANET, IEEE802.11p, WAVE.

## 1. Introduction

In near future, modern vehicles will be equipped with on board intelligent units, which will inform the drivers about a range of safety information, to help them raise their vehicle safety. Apart from safety applications, the commuters will also be able to enjoy the non-safety application e.g. on board internet surfing, multimedia applications and so on. Currently, several ideas around the world are considering vehicular safety applications by means of short range wireless communications. One of the major improvements in vehicular communication is from the Federal Communication Commission (FCC), allocated 75 MHz spectrum at 5.9 GHz for intelligent transport system (ITS) application in 1999 in United States. By adopting the Dedicated Short Range Communication (DSRC) technology, multi-hop ad hoc will become the mainstream technology in modern vehicular environments. In Vehicular Ad-hoc Network (VANET), the vehicles should be able to communicate locally without the need of any centrally managed infrastructure or base stations controlling the medium access. For multi-hop communication, in VANET, data is forwarded to the destination vehicle by using location-based ad hoc routing

protocol instead of IP addresses. There are range of multi-hop routing protocols developed over the years e.g. AODV, DSR, OLR are the few of them. The VANET plays an important role in the development of Vehicular-centered applications where cars collect the local information about the road conditions distribute this information locally and overwhelm local information from the nearby vehicles. Apart from the safety information the non-safety information are also provided to the commuters, for this purpose the Internet Gateways (IGWs) are installed along the roadside to provide a temporary internet access. Though, mobility management is require to handle the mobility of a vehicle in IGW to ensure that the requested data is from the internet always deliver to the appropriate vehicle through IGW. The vehicle must also be able to discover the IGW within VANET even its multi-hop away. To address these problems the IEEE 802.11p task group made some enhancement to the MAC layer for better support of safety and non-safety applications and PHY layer to support communication distance up to 1000m. Also to enhance larger distance, multi-hop communication is to be supported in efficient way. By enforcing such a technology in transportation, congestion problems could be solved out which could save billions of dollars of fuel, also millions of hours of waste of time on the road.

## 2. IEEE 802.11 Standards

The IEEE 802.11 standard describes the PHY and MAC layers specifications of Wireless LANs. There are several methods for data transmission between two nodes at the physical layer. Nodes can use Direct Sequence Spread Spectrum (DSSS), orthogonal frequency distribution modulation (OFDM), or frequency hopping spread spectrum (FHSS). For access mechanism two different methods are used, Distributed Coordination Function (DCF) and Point coordination Function (PCF). The DCF mechanism uses CSMA/CA for access method. In CSMA/CA the exchange of request-to-send (RTS), clear-to-send (CTS), a data and acknowledgement (ACK) are required for each sending data packet. To avoid collision a back-off mechanism is deployed before the start of transmission. If the channel is free, an additional random

time for listening to the channel before starting the transmission is used. This interval is called DIFS (DCF Inter-frame Space). The sender node can start transmission if the channel remains free for DIFS. Contention window (CW) is maintained at each node to determine the amount of time a station should wait before transmitting. The value of CW remains the same each time and ACK is received from the receiver, but it will increase if the transmission fails. The increase in CW will result in an increase of value of random back off timer. Finally, to deal with the problem of hidden node, RTS and CTS are used. An RTS frame is sent by the sender before the transmission and CTS frame is sent back by the receiver to inform his availability for receiving the data. In PCF polling methodology is used which allows the point coordinator node to poll different nodes that need to send the data and to which, if they are polled, they send their packets. Also PCF uses a contention-free period (CFP) and a contention period (CP). During the CFP, a PCF mechanism is used while in CP a DCF method is used.

The two mechanisms used by the WLAN, do not have any room for supporting the real-time traffic since low end-to-end delay or jitter cannot be guaranteed. Polling or in a CSMA/CA point of view, especially in a VANET environment. IEEE is working on IEEE 802.11p extension, which is an enhanced version of IEEE 802.11 standards, designed for VANETs and to support multimedia transmission efficiently in a vehicular environment.

### 3. IEEE 802.11p Enhancements

In this section, we discuss the enhancements related to the IEEE 802.11p standard.

#### 3.1 IEEE 802.11p (WAVE) MAC Enhancements

Most of the changes in IEEE 802.11p standard are related to the MAC layer. MAC layer changes are often software based and can be updated quite easily rather than PHY layer. The enhancements in MAC layer in IEEE 802.11p are listed below.

##### 3.1.1 WAVE Mode

IEEE 802.11 MAC operations are too time intensive, where in a high vehicular environment, vehicular safety communications use cases demand instantaneous data exchange capabilities and cannot afford typical 802.11 method of scanning channels for beacon of BSS and execute multiple handshakes for establishment of communication. It is essential for all IEEE 802.11p compliant devices to have radio configured in a same channel with the same BSSID for safety communication

with no delay. For example if a vehicle crossing another vehicle in the opposite direction, the time for communication may be extremely short due to the vehicle dynamics. WAVE mode is introduced in IEEE 802.11p WAVE for capability enhancement. In WAVE mode a wildcard value is assigned to BSSID for transmitting and receiving of data frames without the need for the node to connect to BSS. This is very beneficial for the vehicle communicating for a short interval of time and for safety communication which do not require additional overhead for simple communication, as long as they use the same channel with wildcard BSSID.

##### 3.1.2 WAVE BSS

The overhead of typical BSS (Basic Service Set) setup is too much expensive for both safety and non safety applications. A vehicle approaching a road side station that offers, suppose services like local information, it can hardly afford few seconds that are required in typical WLAN connection setup, because due to the dynamics of vehicle the total time it stays in the range will be too short then waiting for connection. Analyzing this factor WAVE standard introduced WBSS (WAVE BSS), which is the enhancement of BSS type. In WBSS environment, an STA forms a WBSS by first transmitting an on demand beacon. The WAVE station uses that demand beacon, which uses the well known beacon frame and needs not to be repeated every so often, to advertise a WAVE BSS unlike BSS. Upper layer mechanism above the IEEE 802.11 creates and consumes such advertisements. It contains all the necessary information needed by the receiver station to understand the services offered in the WBSS in order to decide whether to join the WBSS and if needed configure itself into a member of the WBSS. In other words if station decides to join will need only WAVE advertisement for complete joining process with no further overhead.

##### 3.1.3 Wildcard BSSID Usage

The 802.11p WAVE was suggested for safety as a key, the use of wildcard BSSID is supported for stations even they are already belongs to WBSS (i.e. configured with a particular BSSID). Means an STA in WBSS will be still in WAVE mode in order to transmit frames with wildcard BSSID in order to reach its neighbour STAs in cases of safety concerns. Also, an STA already in a WBSS and having its BSSID configured for filtering accordingly can still receive frames from STA's outside the WBSS with wildcard BSSID. The main purpose of BSSID configured with wildcards strengthen the sending and receiving data frames for safety communication but also support signaling of future upper layer protocols in Ad-hoc environment.

### 3.1.4 Distributed Services

The Wave complaint device still support Distributed services. In WAVE BSS the concept of wildcard is used to send and receive data frames, which introduces complications. It is more probable that a radio will be restricted to send a data frame with the wildcard BSSID on if the "To DS" and "From DS" bits are set to 0. Means radios communicating in WAVE BSS environment should send data frames to known BSSID for accessing the DS.

### 3.2 IEEE 802.11p (Wave) PHY Enhancements

In IEEE 802.11p standards, the PHY layer is not change as such because the 802.11a radios are already operate at 5 GHz and it is not difficult to change the configuration of these ratios to work on 5.9 GHz band in U.S and internationally. The purpose of minimum changes to IEEE 802.11 PHY is that WAVE device can communicate efficiently among different fast moving vehicles in the highway environment. On the other hand MAC layer enhancements are basically software updates bases, which is also easy to make, while PHY level enhancement minimal in order to shun designing an entirely new architecture for wireless technology. The few enhancement made are given below.

#### 3.2.1 Channels Frequency

IEEE 802.11p is based on IEEE 802.11a which basically use OFDM modulation scheme for communication with 20 MHz channels, while in IEEE 802.11p 10 MHz channel is used. The implementation of 10 MHz channels is straightforward since it mainly about doubles all of OFDM timing parameters used in the regular 20 MHz 802.11a transmissions. The main reason for this scaling of 802.11a is to describe the increased RMS delay spread in the vehicular environment. For accommodating the large communication range in vehicular environment, four classes of maximum allowable Effective Isotropic Radiation Power (EIRP) up to 30 W (44.8dbm) are allocated in IEEE 802.11p. For emergency vehicles approaching the highest values is reserved which is typically 33dBm for safety relevant messages.

#### 3.2.2 Enhanced Receiver Performance

One of the problems which is well known and is natural property of wireless communications is cross channel interference. In U.S and (expectedly) internationally there are number of channels available for 802.11p deployment

and usage. There is increase concern about cross channel interference among closely distributed vehicle on the road. The measurement presented by (Rai, v., Bai, F., et al.2007).demonstrates the potential for immediate neighbouring vehicles to interfere with each other's communication. If they are using two adjacent channel, for example a vehicle using channel 176 , could be interfered by the vehicle adjacent to it using channel 178, for receiving safety message send by another car ahead. However, IEEE 802.11p introduces some improvement in receiver's performance, required in adjacent channel rejections. There are two categories of requirement listed in the proposed standards. The first category is mandatory and to be understood generally to be reachable with today's chip manufacturers. The second category is more stringent and optional.

## 4. Simulation Setup

The scenarios were built using the IEEE 802.11p standards in the simulators. On the physical layer the most robust PHY mode is chosen (Binary Phase Shift Keying with 50% redundancy, BPSK1/2). The transmission power is 30dBm (1W). Omni-directional antennas are used, so no antenna gain is involved. The center frequency is 5.9GHz with a channel bandwidth of 10MHz, with the 6Mbps data transmit rate.

### 4.1 Scenario Description

The scenario chosen for the evaluation is the highway scenario. In the first scenario a single hop communication between the mobile node and the roadside server is represented. The first scenario consist three different speed limits for the mobile nodes. First the node move with the 32 km/h (20mi/h), then the same node with the same environment and setting only the speed is alter to 65 km/hr (40mi/h), and 97 km (60mi/h). The second scenario is the same as the first one but a second server is planted to check the mobile node communication with that server multi-hop away. In the third scenario we have two mobile nodes in the same direction. The first node is moving with 32km/h (20mi/h), while the second node's speed changes. At first the second node move with 32km/h(20mi/h), then 65 km/h(40mi/h) and 97km/h(60mi/h) respectively. The communications between the nodes are tested using the FTP services. The IEEE 802.11p protocol stack, the channel and mobility model were implemented in our discrete event simulator OPNET™ modeler.

## 5. Simulation Results

In this section, we present the simulation results in different configuration scenarios.

### 5.1 Single Hop Communication

The simulation results are presented next in Figures 1-2. The simulation for single hop communication consists of three simulations. The moments of car was 32 km/h (20mi/h), 65km/h (40mi/h) and 97km/h (60 mi/h) respectively. The graph represents the comparative average wireless through and average delay put in figures 1-2 of nodes in three different speed ranges.

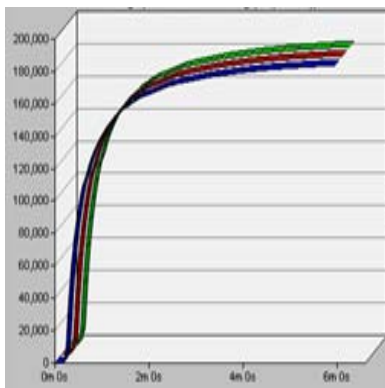


Figure 1: Average through put

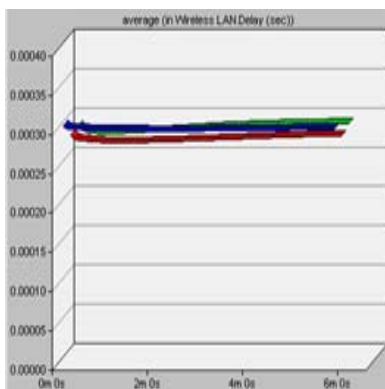


Figure 2: WLAN delay (sec)

AODV routing protocol is used for the routing purpose. The average delay recorded of the route discovery is presented in figure 3. As one can see it is the same in different speed scenario. Also for signaling purposes the H323 protocol is used for registration, call admission control, and call signaling. In figure 4 the call admission and registration time is shown in three dimensions graph to so the impact of the vehicle density and speed on each of the simulated matrices can be easily evaluated.

In next graphs shown in figure 5-6, the performance of AODV routing protocols is shown. The average packets send and received in all different speed scenarios from sender and receiver are about the same.

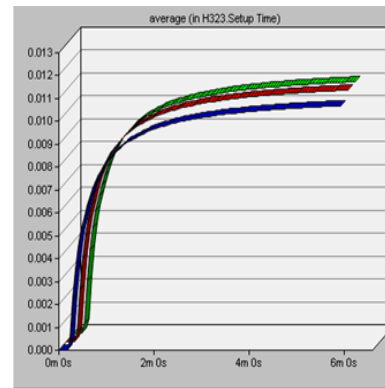


Figure 3: AODV route discovery time

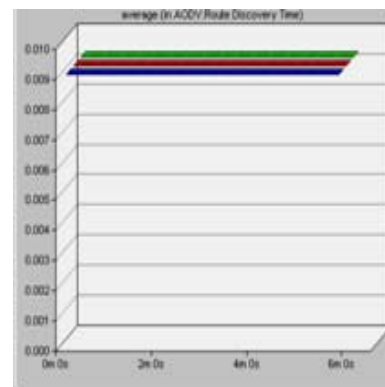


Figure 4: H323 call setup time

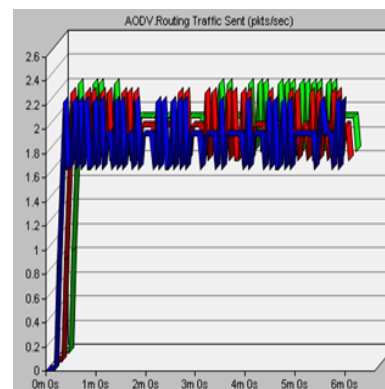


Figure 5: AODV routing traffic sent

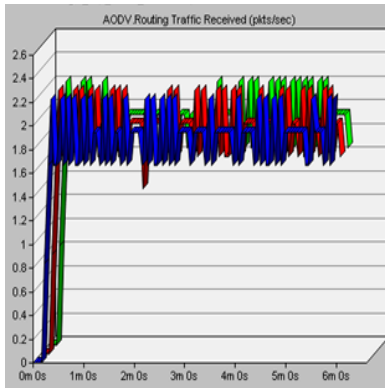


Figure 6: AODV routing traffic received

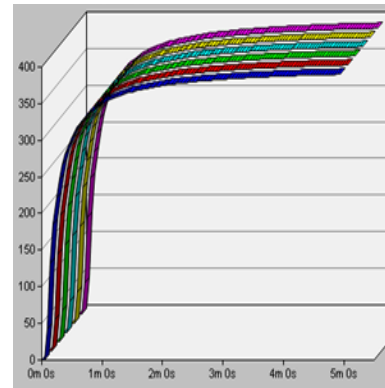


Figure 8: Packets transmission

### 5.2 Second Scenario Multiple Hops

In the second scenario, the whole system is kept the same except a new fixed server is introduced to extend the limit of moving node. The second server is kept in such distance that it is not interrupted with the channels of the first server. The second server has the same properties as first. In this evaluation, we tried to communicate the moving node with the server a hope away. The figures 7-8 shows the throughput and voice packets send/receive among nodes and servers. Figure 9-10 represents the AODV route discovery and H323 call setup time.

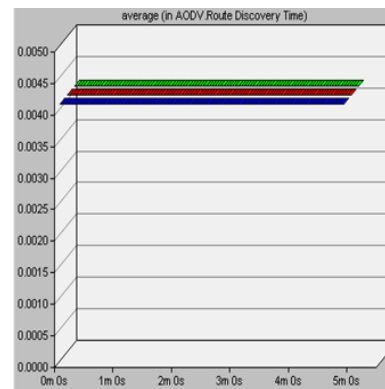


Figure 9: AODV route discovery time

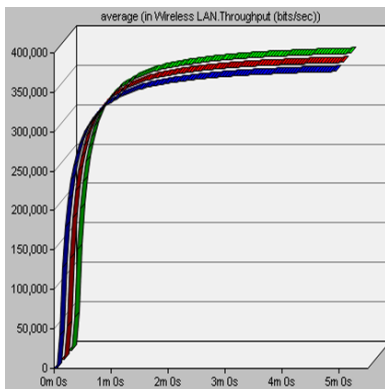


Figure 7: Wireless LAN throughput

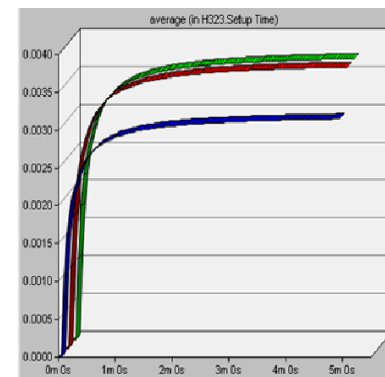


Figure 10: H323 call set up time

Figure 11 represent the average end-to-end delay of the voice packet send across the multi-hop. The multi-hop communication is carried out with the server not in the direct range of the moving range. This is about 0.06 seconds. While WLAN delay is about 0.0003seconds represented by figure 12.

Figure 13-14 represent the AODV routing traffic (packet/sec) sent and received. The sent packets are about

three packets per second and receive traffic about six packets per second. It shows the amounts of received packets are from both the servers. In this graph the mobile node efficiently communicating with the server not in the direct range

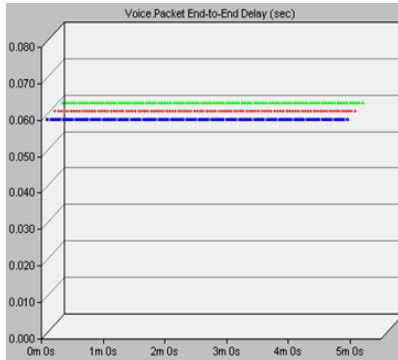


Figure 11: End-to-end voice packet delay

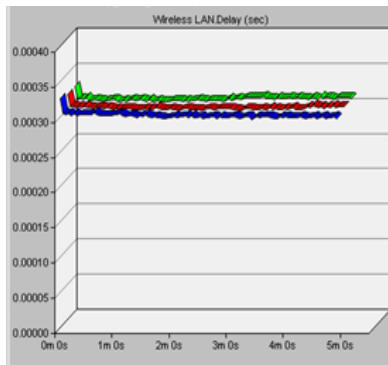


Figure 12: Average WLAN delay

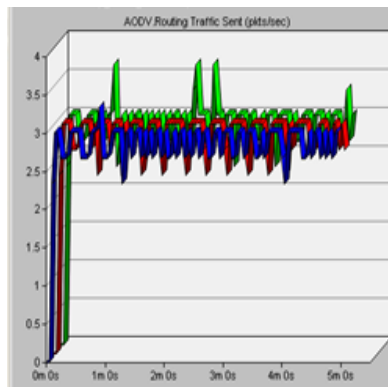


Figure 13: AODV routing packet sent

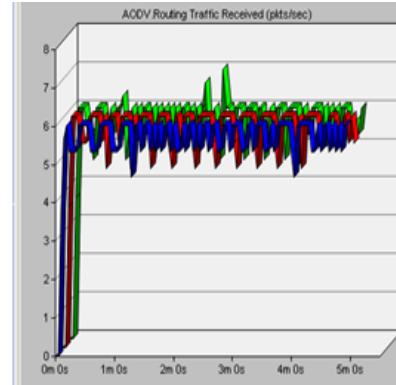


Figure 14: AODV routing packets received

### 5.3 Mobile Node to Node Communication

The third scenario is based on node to node communication while on both the nodes are on the move. This kind of communication is typical Ad-hoc communication, as it is direct communication between the communicating nodes. A FTP file transfer is chosen for file upload or download operation. Figure 15-16 represent the Wireless LAN through put and average delay occur between the moving nodes. The scenario where both cars moving with 32 km/h the average wireless delay is low because of the equal speed, then the nodes moving with 65 km/h and 97 km/h respectively. In fig 16 the delay variation fluctuates so often because of the dynamics of the moving vehicle, but as a whole the delay is still very low and affordable in vehicular environment.

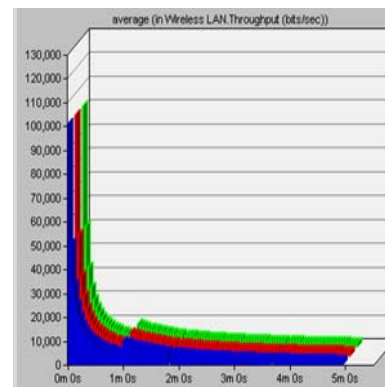


Figure 15: Average WLAN throughput

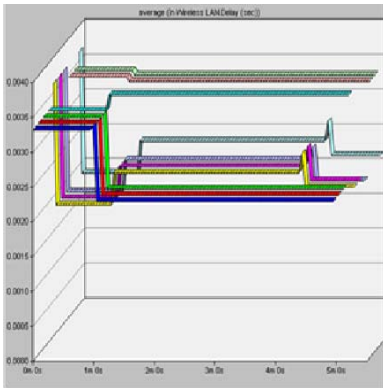


Figure 16: Average WLAN delay

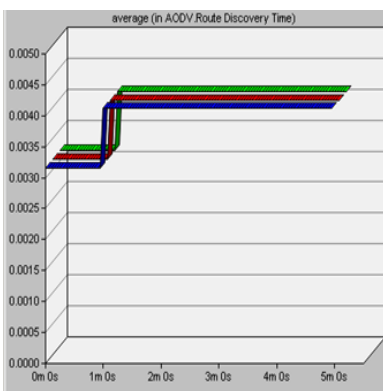


Figure 17: AODV route discovery

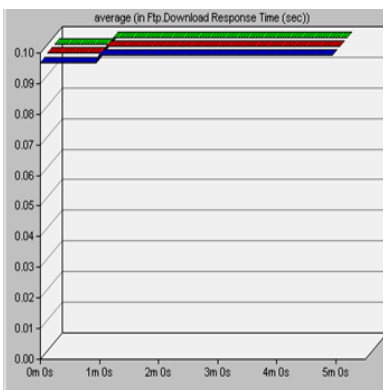


Figure 18: FTP downloads response

Figure 17 represent the average AODV routing protocol route discovery. Figure 18 represents the Average response from the FTP server to download or upload an FTP file.

## 6. Conclusion

In this paper, the enhancement of IEEE 802.11p (WAVE) is discussed. To investigate its impact on the single hop and multi-hop communication in high vehicular environment, real time voice applications are used. AODV is used as a multi-hop routing protocol and H232 codec for voice applications. Different parameters were checked during the simulation regarding the quality of communication. To investigate the communication between the neighboring nodes, FTP services are used. The results show that WAVE complaint applications and devices can greatly improve the communication range and performance of VANET, by supporting efficient multi-hop communication and reducing delay and connection time.

## References

- [1] Ko Y F, Sim M L Nekovee M. (2006), "Wi-Fi based broadband and wireless access for user on the road", *BT Technology Journal*, Vol 24, No. 2, pp 123-129
- [2] Berger, I. (2007) "Standards for Car Talk", *The Institute*, Vol 24, No.1, pp 1-6
- [3] Performance Report (Accessed 20 July 2008). "<http://www.its.pdx.edu/project.php?id=2007-20>".
- [4] Jiang, D. and Delgrossi, L. (2008), "IEEE 802.11p: Towards an International Standard for Wireless Access in Vehicular Environment", *Vehicular Technology Conference, 2008. VTC Spring 2008. IEEE*, pp. 2036-2040.
- [5] Palazzi, C. E., et al. (2007). "Facilitating Real-Time Application in VANETs through Fast Address Auto-Configuration", *Consumer Communications and Networking conference, 2007.CCNC 2007 4th IEEE*: pp.981-985.
- [6] Vaqar, S. A. and Basir, O. (2007), "Smart Protocol for Communication in Mobile Ad Hoc Network of Vehicles",

Telecommunications, 2007.ITST '07.7th International Conference on ITS. Sophia Antipolis, IEEE: pp.1-6.

[7] IEEE 1609.0 (2007), "Trial Use Standard for Trial Use Standard for Wireless Access in Vehicular Environments (WAVE) - Architecture or Wireless Access in Vehicular Environments (WAVE) - Architecture", IEEE P1609.0/D02

[8] IEEE 1609.4 (2006), "IEEE Trial-Use Standard for Wireless Access in Vehicular environments (WAVE) - Multi-channel Operation", pp.c1-74.

[9] ITS Website (2007), "US Department of Transportation", <http://www.its.dot.gov/index.htm>, (Access 25 July)

[10] Royer, E. M. and Perkins, C. E. (2000), "An implementation Study of the AODV routing Protocol", Wireless Communication and Networking Conference, 2000, WCNC, 2000 IEEE. Chicago, IL, USA. vol. 3: pp. 1003-1008

[11] Stibor, L. and Zang, Y. (2007), "Neighborhood evaluation of vehicular ad-hoc network using IEEE 802.11p", Proceedings of the 8th European Wireless Conference, Paris, France.

[12] FCC, (2006). "FCC Report and Order 06-110: Amendment of the Commission's Rules Regarding Dedicated Short-Range Communication's Services in the 5.850-5.925 GHz Band".

[13] IEEE (2007), "Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications", IEEE Std. 802.11, 2007.

[14] IEEE (July 2007), "IEEE P802.11p/D3.0, Draft Amendment for Wireless Access in Vehicular Environment (WAVE)".

[15] ASTM (2002), "Standard Specification for Telecommunications and Information Exchange Between Roadside and Vehicle Systems — 5 GHz Band Dedicated Short Range Communications (DSRC) Medium Access Control (MAC) and Physical Layer (PHY) Specifications", STD E2313 - 02.



# Video Authentication: Issues and Challenges

Saurabh Upadhyay<sup>\*</sup>, Sanjay Kumar Singh<sup>†</sup>

<sup>\*</sup>Department of Computer Science & Engineering, SIT, Gujarat-India

<sup>†</sup>Department of Computer Engineering, IT-BHU, Varanasi-India

## Abstract

Video authentication aims to ensure the trustworthiness of the video by verifying the integrity and source of video data. It has gained much attention in the recent years. In this paper we present the issues in the designing of a video authentication system. These issues include the classification of tampering attacks, levels of tampering attack and robustness. Further we present the categorization of existing video authentication techniques with their shortcomings. Moreover we have also given the challenging scenarios in which the video authentication would be a critical task.

**Keywords:** Video Authentication, Fragile Watermarking, Digital Signature, Intelligent Techniques, Tampering Attacks

## 1. Introduction

With the rapid innovation and development in digital technologies, video applications are infiltrating into our daily lives in breakneck speed from traditional television broadcasting to modern vulnerable communication media such as Internet/Intranet, wireless communication and consumer products such as VCD/DVD. In some applications the authenticity of video data is of paramount interest such as in video surveillance, forensic investigations, law enforcement and content ownership [33]. For example, in court of law, it is important to establish the trustworthiness of any video that is used as evidence. As in another scenario, for example, suppose a stationary video recorder for surveillance purpose, is positioned on the pillar of a railway platform to survey every activity on that platform along a side, it would be fairly simple to remove a certain activity, people or even an event by simply removing a handful of frames from this type of video sequences. On the other hand it would also be feasible to insert, into this video, certain objects and people, taken from different cameras and in different time. So video authentication is a process which ascertains that the content in a given video is authentic and exactly same as when captured.

### 1.1 Motivation behind Video Authentication

A video clip can be doctored in a specific way to defame an individual. In the recent years, several cases have been reported where the eminent personalities of the society were caught in illegal activities in the video recordings made by so called journalists. However in the absence of foolproof techniques to authenticate the video it is difficult

to trust on such reports. On the other hand criminals get free from being punished because the video (used as evidence), showing their crime cannot be proved conclusively in the court of law. In the case of surveillance systems, it is difficult to assure that the digital video produced as evidence, is the same as it was actually shot by camera. In another scenario, a news maker cannot prove that the video played by a news channel is trustworthy; while a video viewer who receives the video through a communication channel cannot ensure that video being viewed is really the one that was transmitted [6]. In the scenario of sensitive cases where a video is produced as a witness in the court of law, even a small modification may not be acceptable. However there are some scenarios where editing also may be allowed while keeping intact the authenticity of the video. For example after shooting the video, a journalist may need to perform some editing before broadcasting it on a news channel. In such a case a video authentication system should be able to allow editing on the video up to a certain level ensuring the authenticity of the video [38]. These are the instances where malicious modifications cannot be tolerated. Therefore there is a compelling need for video authentication

Although traditional data authentication technology for message integrity was mature, video authentication is still in its early development stage and many fundamental questions remain open [28]. For example, for a number of different authentication algorithm developed over the past few years, it is difficult to affirm which approach seems most suitable for ensuring the integrity adapted to videos [28]. There is a need for synthesizing literature to understand the nature of the problem, identify the potential for research issues, standardize new research area and evaluate the relative performances of different approaches. The aim of this paper is to examine the status and issues of video authentication techniques and to assess their strengths and weakness in the reference of different tampering attacks. This paper is organized as follows. In Section 1 the notion of video authentication and framework are briefly introduced followed by a discussion of motivation behind video authentication. Section 2 explains the issue of robustness for video authentication. Security related issues are discussed in detail in section 3. Section 4 provides a concise review of existing techniques for video authentication. Some of the new challenging scenarios are

briefly introduced in section 5. Finally the summary and future research directions are discussed

## 2. Robustness

Any video applications may have at least three parties: Producer, receiver and the third party. The producer generates the video and the receiver receives the video from producer via third party. Here the third party is a general and wide concept. It could be either a storage device in consumer products (such as CD/DVD) or a busy and noisy channel in video transmission. Further a receiver can also be a third party if, after receiving the video, it forwards the video to any other party. The malicious attacker targets this third party category for altering the video content. Video authentication is the process which ascertains that the content in a given video is authentic and exactly same as when captured. Lin and Chang [8] classified the multimedia authentication techniques into two categories: Complete authentication and content authentication. The techniques which are proposed for complete authentication consider that the multimedia data, which have to be authenticated, have to be exactly the same as the original one. No change in the multimedia data is allowed. In content authentication, as long as the meaning of multimedia data remains unchanged, the received multimedia data is considered as authentic, regardless of the processing or transformation the multimedia data has undergone. Of course, video authentication should be content authentication because a receiver must not obtain an exact copy of the original video without any distortion, necessarily. For instance, due to its bigger size in storage, digital videos are usually compressed and most video compression, such as MPEG 1/2/4 are lossy compression. And definitely the de-compressed video is not identical to the original one. However, it should still be considered to be authentic. Another example is video transcoding in which the bit rate of a video stream is adjusted to adapt to variable transmission channel.

Thus a video authentication system theoretically should be robust enough to discriminate all normal video processing operations from malicious tampering attack. A robust video authentication system should tolerate the incidental distortion, which may be introduced by normal video processing such as compression, resolution conversion and geometric transformation, while being capable of detecting the intentional distortion, which may be introduced by malicious attack. However, it is a difficult task to define all acceptable video processing operations due to the huge diversity of video applications. For example, the object based video processing operations such as rotation, scale and translation (RST) is very different from the traditional frame-based video processing operations. The video authentication system should also be sensitive to malicious manipulations.

## 3. Security issues of video authentication.

A continuous video sequence  $V_c(x, y, t)$  is a scalar real valued function of two spatial dimensions  $x$  and  $y$  and time

$t$ , usually observed in a rectangular spatial window  $W$  over some time interval  $T$ . If  $B(x, y, t)$  is modification vector then the tampered video  $M_c(x, y, t)$  would also be a scalar real valued function of spatial dimensions  $x$  and  $y$  and time  $t$  as follows:

$$M_c(x, y, t) = B(x, y, t) + V_c(x, y, t)$$

When the content of information, being produced by a given video sequence is maliciously altered, then it is called tampering of video data. It can be done for several purposes, for instance to manipulate the integrity of an individual. Since a wide range of sophisticated and low cost video editing software are available in the market that makes it easy to manipulate the video content information maliciously, it projects serious challenges to researchers to be solved.

### 3.1. Video Tampering Attacks

There are several possible attacks that can be applied to alter the contents of a video data. Formally a wide range of authentication techniques have been proposed in the literature but most of them have been primarily focused on still images. In several applications, due to large availability of information in video sequences, it may be more significant if the authentication system can tell where the modifications happened (It indicates the locality property of authentication) and how the video is tampered [5]. On considering these where and how, the video tampering attacks can have different classifications. A lot of works have been done that briefly address the classification based on where [33], [5]. And some papers address the classification based on how [34]. In general, finding where the multimedia data is altered is more efficient than to find out how the multimedia data is tampered. When a video is being recorded by a video recording device, it captures the scene which is in front of the camera lens, frame by frame, with respect to time. Number of frames being captured by video recording device in a second depends on the hardware specification of the device. Thus a video sequence can be viewed as a collection of consecutive frames with temporal dependency, in a three dimensional plane. This is called the regional property of the video sequences. When a malicious alteration is performed on a video sequence, it either attacks on the contents of the video (i.e. visual information presented by the frames of the video), or attacks on the temporal dependency between the frames. Therefore based on the regional property of the video sequences, we can broadly classify the video tampering attacks into three categories: spatial tampering attacks, temporal tampering attacks and the combination of these two, spatio-temporal tampering attacks [5]. They can be further classified into their subcategories.

#### 3.1.1. Spatial Tampering

In spatial tampering malicious alterations are performed on the content of the frames ( $X$ - $Y$  axis). The operations that can be done as tampering attack in spatial tampering are cropping and replacement, morphing, content (object)

adding and removing etc [5]. These attacks can be efficiently performed with the help of video editing software as *Photoshop*, etc.

### 3.1.2. Temporal Tampering

In temporal tampering manipulation is performed on the sequence of frames. The focus is on the temporal dependency. Temporal tampering attacks are mainly affecting the time sequence of visual information, captured by video recording devices. The common attacks in temporal tampering are frame addition, frame removal and frame reordering or shuffling.

### 3.1.3. Spatio-Temporal Tampering

Spatio-temporal tampering attacks are the combination of the both kinds of tampering attacks. Frame sequences are altered as well as visual contents of the frames are modified in the same video. The authentication system should be able to identify both kinds of tampering.

All these tampering are further classified into their subcategories. Spatial tampering can be in effect either at block level or at pixel level. In both the cases the objects of the frames of the video are altered. Further the objects of the frames are classified into two categories: Foreground objects and Background objects. The foreground objects are those which are captured as individual elements, excluding the background, in a frame. And the background object is the background part of the frame excluding all of the foreground objects. The different pieces of visual information shown in the frames of the video are altered in spatial tampering. Basically the contents of the video frames are treated as objects. Based on these objects and their classification the spatial tampering can be further classified as following figure shows.

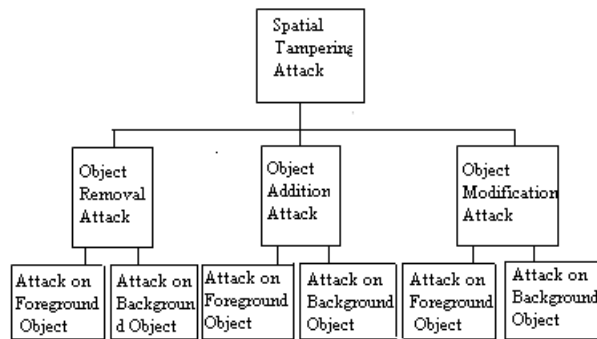


Fig.1. Spatial Tampering Classification

Fig. 1 shows an explicit classification of spatial tampering attacks in reference of objects of the frames.

#### 3.1.1.1. Object Removal Attack

In object removal attack, the objects of the frames of the video are eliminated. This kind of attack is commonly performed where a particular person wants to hide his/her presence in a certain sequence of frames. With this kind of attack he /she may disappear in a specific time domain,

recorded in the video. This attack can be performed with both kinds of object, foreground objects and background object, as shown in Fig. 2

#### 3.1.1.2. Object Addition Attack

When an object is inserted in a frame or in a set of frames then there is a kind of spatial tampering attack: say Object addition attack. In any video sequence which can be treated as evidence, an additional object can be pasted in a frame or set of frames, with the help of sophisticated video editing software to mislead the investigation agencies as well as court of law. As shown in fig. 3, it can also be performed with both kinds of objects, foreground objects and background objects.

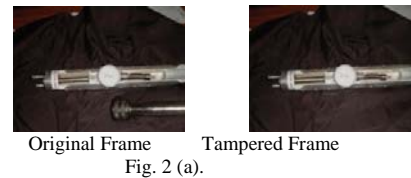


Fig. 2 (a).



Fig. 2 (b).

Fig.2. Example of object removal attack. Fig.2 (a) shows object removal attack with foreground object, where a small device is removed from the original frame in tampered frame. Whereas Fig.2 (b) shows the object removal attack with background object. Here a small object on the right side of the wall is eliminated from the original frame in tampered frame.

#### 3.1.1.3. Object Modification Attack

In Object modification attack, an existing object of the frame(s) can be modified in such a way that the original identity of that object is lost, and a new object may be in appearance which is totally different from the original object. The object modification attacks can be existed in many prospects in the given video. For instance, the size and shape of the existing object may be changed, the colour of the object may be changed or it may be discoloured, and with the help of additional effect the nature of the object and it's relation with other objects also may be changed. In fact it is very hard to detect this kind of attack for authentication systems, since these attacks are performed at pixel level. The authentication systems should be robust enough to differentiate this kind of attack with the normal video processing operations. Fig.4 shows a typical example of object modification attack where the face of a person has been changed in such a way that a new person's face is introduced in the altered frame. These attacks can also be performed with both kinds of objects, foreground and background objects.

Besides spatial tampering, temporal tampering attacks have also sub classifications. Temporal tampering attacks can be performed at scene level, shot level and frame level, but the

primary focus is on attacking the temporal dependency of the frames of the video.

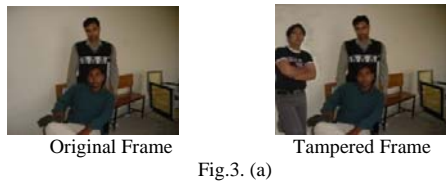


Fig.3. (a)

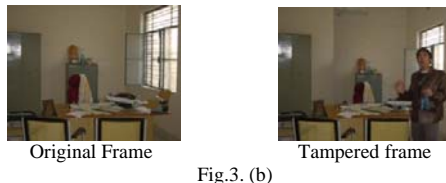


Fig.3. (b)

Fig.3. Example of Object Addition attack. In original frame of Fig.3 (a) two persons are there as major foreground objects, while in tampered frame of Fig.3 (a) an additional person as a foreground object is added. In tampered frame of the Fig.3 (b), not only a foreground object is added but also an additional wall as a background object, in the middle of the frame, is added.



Fig.4. Example of Object modification attack. The face of the person in original frame is modified in tampered frame, in such a way that the new face of the person cannot be identified as the same as in original frame.

We call it ‘Third dimensional (dimension with respect to time) attack’ on the video sequences. Therefore based on this third dimensional attack we can classify the temporal tampering attacks into following categories.

### 3.1.2.1. Frame Addition Attack

In frame addition attack, additional frames from another video, which has the same statistical properties, are intentionally inserted at some random locations in a given video. This attack is intended to camouflage the actual content and provide incorrect information [33]. A typical example of the frame addition attack is shown in fig. 5.

### 3.1.2.2. Frame Removal Attack

In frame removal attack the frames of the given video are intentionally eliminated. In this kind of attack frames or set of frames can be removed from a specific location to a fixed location or can be removed from different locations. It depends upon the intention. Commonly this kind of tampering attack is performed on surveillance video where an intruder wants to remove his/her presence at all. Fig. 6 shows a typical example of frame removal attack

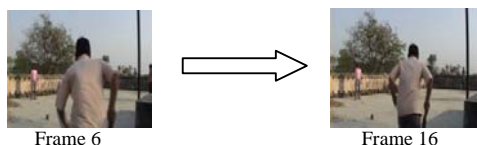


Fig.5 Example of Frame addition attack. In first row the original frame sequence from frame 6 to frame 16 has been shown. After attack, the second row of the frames shows the altered frame sequence in which a new frame is inserted between frame 6 and frame 16. And frame 16 becomes frame 17.

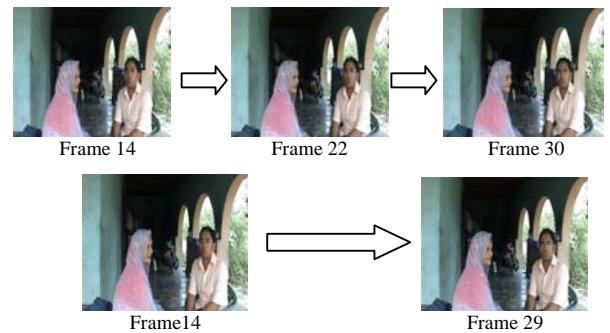


Fig.6 Example of Frame removal attack. The first row of this figure shows the original frame sequence with frame 14, frame 22 and frame 30. In second row of the frame sequence, which shows the tampered frame sequence with frame removal attack, frame 22 is eliminated from the video and hence frame 30 becomes frame 29.

### 3.1.2.3. Frame Shuffling Attack

In frame shuffling attack, frames of a given video are shuffled or reordered in such a way that the correct frame sequence is intermingled and wrong information is produced by the video as compared to original recorded video. Fig. 7 shows a typical example of frame shuffling attack where two frames are shuffled.

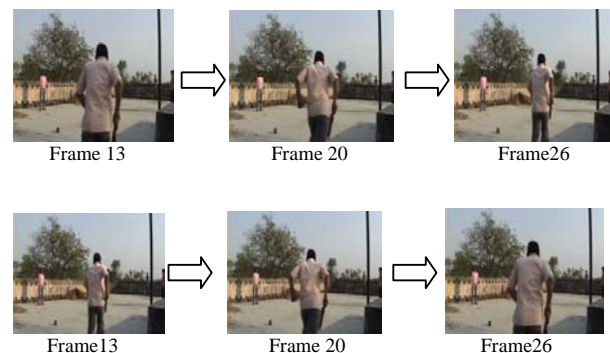


Fig.7. Example of Frame shuffling attack. The first row of this figure shows the original frame sequence with frame 13, frame 20 and frame 26. After the frame shuffling attack, the original frame sequence is tampered as shown in second row of the figure where the positions of frame 13 and frame 26 have been changed.

## 3.2. Levels of Tampering Attacks

In addition of these types of tampering attacks, tampering can be done at different levels in video sequences.

### 3.2.1. Scene Level

When the tampering attacks are performed at scene level then a whole scene of the video sequence is manipulated in such a way that, not even the scene itself is modified but also in the reference of the given video, the scene of that video is modified. It means spatial and temporal both kinds of tampering can be done at scene level.

### 3.2.2. Shot Level

In shot level tampering any particular shot of the given video is modified in reference to the given video. In shot level tampering a shot can be added or removed from the video. It can also be performed with all kinds of tampering attacks.

### 3.2.3. Frame Level

When frames of the given video are maliciously modified, then it is called tampering at frame level. Frame removal, frame insertion and frame shuffling are the common tampering attacks that can be performed at frame level. In other words, temporal tampering attacks are commonly performed at frame level.

### 3.2.4. Block level

In block level tampering, tampering attacks are performed on the blocks of the video frames. The content of the video frames are treated as blocks on which the tampering attacks are applied. Blocks (a specified area on the frame of the video) can be cropped and replaced, morphed or modified in any way in block level tampering. Spatial tampering attacks are commonly performed at block level.

### 3.2.5. Pixel level

In pixel level tampering contents of the video frames are modified at pixel level. This is the smallest level in video sequences at which tampering attacks can be performed. The video authentication system should be robust enough to differentiate the normal video processing operation and pixel level tampering, since many normal video processing operations are performed at pixel level. Spatial tampering attacks are commonly performed at pixel level. All these levels of tampering show the different aspects of tampering.

## 4. State of the art review

By definition, authenticity means sometimes “as being in accordance with fact, as being true in substance”, or “as being what it professes in origin or authorship, as being genuine” [30]. Another definition of authentication is to prove that something is “actually coming from the alleged source or origin” [31]. Video authentication, in general has received considerable attention by academia and practitioners over the last few years.

A typical video authentication system is shown in fig. 8. For a given video, authentication process starts with feature

extraction. After that, with a specific video authentication algorithm, the authentication data  $H$  is generated using the features  $f$  of the video. This authentication data  $H$  is encrypted and packaged with the video as a signature or alternatively it can be embedded into the video content as a watermark. The video integrity is verified by computing new authentication data  $H'$  for the given video. The new authentication data  $H'$  is compared with decrypted original authentication data  $H$ . If both are matched, the video is treated as authentic else it is constructed to be tampered.

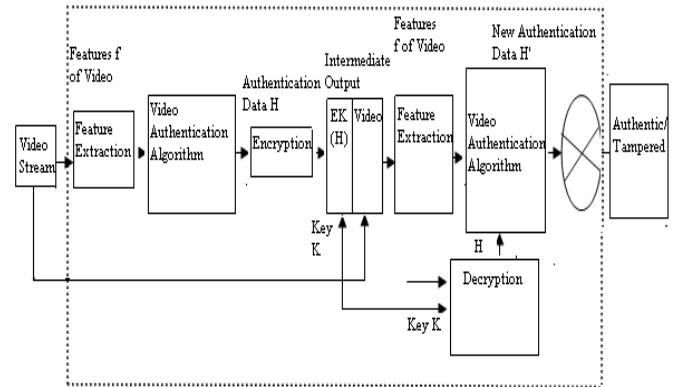


Fig. 8 A typical video authentication system.

## 4.1. Classification of Authentication Techniques

In past few years, watermarking and digital signatures have been widely used for the purpose of video authentication. Different techniques have their own advantages and shortcomings. Fig. 9 represents the tree structure of techniques which have been commonly proposed for the purpose of video/image authentication. In fact fragile watermarking and digital signatures are the two basic schemes for authentication [5]. Moreover there has also been worked on intelligent techniques for video authentication. Apart from these digital signature, fragile watermarking and intelligent techniques, some other authentication techniques are also introduced by researchers. We are giving here a brief classification of video authentication techniques.

### 4.1.1. Digital Signature

Integrity of multimedia data can be greatly verified by digital signature. For the authentication of multimedia data, it was first introduced by Diffie and Hellman in 1976[26]. For the purpose of authentication, digital signatures can be saved in two different ways. Either they can be saved in the header of the compressed source data, or it can be saved as an independent file. Further they can be produced for verification. In the prospective of robustness, since the digital signature remains unchanged when the pixel values of the images/videos are changed, they provide better

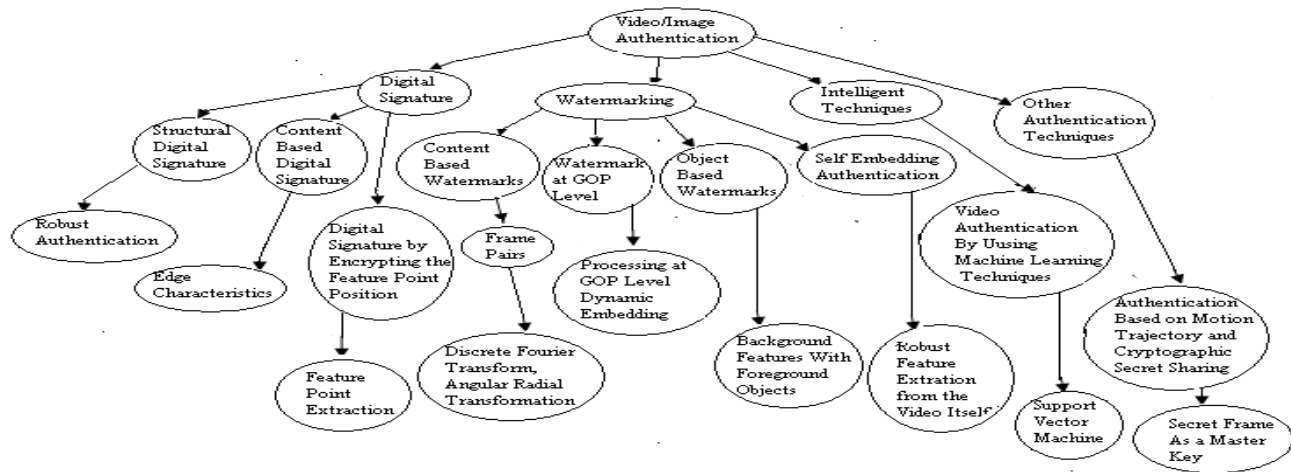


Fig. 9 Tree Structure of Authentication Technique

results. In the digital signature authentication, the digital signature of the signer to the data depends on the content of data on some secret information which is only known to signer [27]. Hence, the digital signature cannot be forged, and the end user can verify a received multimedia data by examining whether the contents of data match the information conveyed in the digital signature. Ching-Yung et al [8] proposed a scheme in which two types of robust digital signatures are used for video authentication in different kinds of situations. The first type of authentication signature is used in situation where the *GOP* (Group of Pictures) structure of the video is not modified, after transcoding or editing processes. The situation, where the *GOP* structure is modified and only the pixel values of picture will be preserved; a second type of digital signature is used. In another work, video authentication is done by generating digital signatures for image blocks and using them as watermarks [3]. In this approach localization packet, watermark insertion is done via *LSB* modification of pixel values. As compared to [2] where video tampering is identified through an analysis of watermark sequencing, here (explicit) block ID's are used for this purpose. The Johns Hopkins University Applied Physics Laboratory (APL) has developed a system for authentication of digital video [7]. The authentication system computes secure computer generated digital signatures for information recorded by a standard digital video camcorder. While recording, compressed digital video is simultaneously written to digital tape in the camcorder and broadcast from the camera into the Digital video authenticator. In this authentication system, video is separated into individual frames and three unique digital signatures are generated per frame-one each for video, audio and (camcorder) control data- at the camcorder frame rate. Here the key cryptography is used. One key, called a "private" key is used to generate the signatures and is destroyed when the recording is complete. The second a "public" key is used for verification. The signatures that are generated make it easy to recognize tampering. If a frame has been added it

would not have a signature and will be instantly detected and if an original frame is tampered the signature would not match the new data and it will be detected in verification process. Ditmann [17] and Queluz [18] used the edge /corner of the image as the feature to generate the digital signature. They claimed this feature is robust against high quality compression and scaling but the problem is that the signature generated based on the edge is too long, and the consistency of the edge itself is also a problem. The digital signature and watermarking, are able to detect regions that have been tampered, but often they are too fragile to resist incidental manipulations. For this type of incidental manipulations structural digital signature [23] can be used for image authentication. This approach makes use of an image's content to construct a structural digital signature (SDS) for image authentication. In this approach [23], many incidental manipulations which can be detected as malicious modifications in other digital signature verifications or fragile watermarking schemes, can be ignored. In the scenario of a station streaming video over network, it is significant for the audiences to have guarantees that the video stream they are watching is indeed from the station. Schemes that are used for this purpose can prevent the malicious parties from injecting commercials or offensive materials into the video streams. Actually this problem has been covered in information security called streaming signing [12] [13], which is an extension from message signing by digital signature schemes. A separate authentication code is written in [37] from the blocks of the video frames. Here the authors Po-Chyi Su et al use the approach of scalar/vector quantization on the reliable features. Once the authentication code is written, it is transmitted along with the video. Thus the authenticity of the given video content can be checked by matching the extracted feature with the transmitted authentication code. Navajit Saikia and Prabin K. Bora present a scheme for video authentication in [15] that generates the message authentication code (*MAC*) for a group of frames (*GOF*) using coefficients from the last but

one high pass band at full level of temporal wavelet decomposition. This digital signature based scheme uses temporal wavelet transform for the generation of message authentication code. After the extraction of *GOFs* from the video, these *GOFs* are recursively decomposed into high pass band up to a certain level using temporal wavelet transform. At this level the high pass band consists of two frames. In the signature generation process, these frames are divided into some blocks of fixed sizes. These blocks are randomly mapped on to a set of groups, using a mapping key in such a way that each group contains equal number of blocks. With the transform coefficients and these groups of blocks, a set of linear combination values is evaluated for each frame in the high pass band. And with these sets of linear combination values, message authentication code (*MAC*) is obtained for the *GOF*. In the signature verification process, the distances  $d(MAC_i, 1, MAC^i, 1)$  and  $d(MAC_i, 2, MAC^i, 2)$  are calculated where  $d$  is any distance measure and  $\{MAC_i, 1, MAC_i, 2\}$  is the *MAC* of  $i^{th}$  *GOF* of the original video and  $\{MAC^i, 1, MAC^i, 2\}$  is the *MAC* of corresponding *GOF* calculated at receiver site. Here the *GOF* of the video would be authentic if these two distances are below some predefined threshold values, otherwise tampered. This authentication scheme would be advantageous for spatio-temporal manipulations, since it is effective for spatial tampering as well as for temporal tampering. Similar to Dittmann's [17] content based digital signature approach for image/ video authentication using edge characteristics, Bhattacharjee and Kutter [25] proposed a scheme to generate a digital signature by encrypting the feature points positions in an image. In this approach authentication is accomplished by comparing the positions of the feature point extracted from the targeted image with those decrypted from the previously encrypted digital signature.

#### 4.1.2. Watermarking

Watermarking always remains a significant issue for solving authentication problems regarding digital multimedia data, in past few years. A wide variety of watermarking techniques have been proposed by various researchers in literature. Based on the application areas, watermarking can be classified in different categories [34]. Beside to ensure the integrity of the digital data and recognizing the malicious manipulations, watermarking can be used for the authentication of the author or producer of the content. Watermarks can be embedded with the multimedia data, without changing the meaning of the content of the data. The advantageous feature with the watermarks is that, they can be embedded without degrading the quality of multimedia data too much. Since the watermarks are embedded in the content of video data, once the data is manipulated, these watermarks will also be modified such that the authentication system can examine them to verify the integrity of data. In [4], authors describe the use of video authentication template, which uses a bubble random sampling approach applied for

synchronization and content verification in the context of video watermarking. The authentication template is introduced in order to ensure temporal synchronization and to prevent content tampering in video sequences [4]. Basically in past few years, an increasing use of digital information in our society and availability of very sophisticated and low cost video editing software creates problems associated with copyright protection and authentication. The owners or producers of information resources are being worried of releasing proprietary information to an environment that appears to be lacking in security [9]. On the other hand with the help of powerful video editing software one can challenge the trustworthiness of digital information. In [9], M. P. Queluz presents the generic models with labelling and watermarking approaches for content authentication. In labelling based approach authentication data are written in separate file [9], while in watermarking based approach the authentication information is embedded in the frames. In this labelling-based authentication system, features *C* and *C'* are extracted from the original and modified pictures respectively as according:

$$C = f_c(I) , C' = f_c(\hat{I})$$

In order to assure the authenticity of the label content, it is signed in a trustworthy way, that is, the label is encrypted with a private key ( $K_{pr}$ ). The label content is produced as:

$$L = EK_{pr}(C, C_l)$$

Where  $C_l$  is optional information, say *Complementary Information*, about the frame and its author, assigned by an author society. In the authentication system the corresponding public key  $K_{pu}$  is used to decrypt the label, producing:

$$C, C_l = EK_{pu}(L)$$

Moreover in [9] M. P. Queluz presents two classical image features for image/video content authentication. The first image feature is concerned with second order image moments. The second feature relies on image edges and it takes the problem of image/video authentication from a semantic view [9]. In image moments feature, for a two dimensional continuous function  $f(x, y)$ , the moments of order  $(p + q)$  is defined as

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x^p y^q f(x, y) dx dy$$

For  $p, q = 0, 1, 2, \dots \dots \dots$

For a digital image the above equation would be as follows:

$$m_{pq} = \sum_i \sum_j i^p j^q f(i, j)$$

Where  $f(i, j)$  represents image color values at pixel site  $(i, j)$ . Moments are usually normalized dividing it by the image total mass, defined as  $\sum_i \sum_j f(i, j)$ . Chang-yin Liang, et al [16] proposed a video authentication system which is robust enough to separate the malicious attack from natural video processing operations with the cloud watermark. In the authentication system [16], first of all, the video sequence is split into shots and the feature vectors are extracted from each shot. Then the extracted feature is used to generate watermark cloud drops with a cloud generator

[16]. Here, for robustness, a content based and semi fragile watermark is used for authentication. In this authentication technique DCT coefficients are evaluated first by partially decoding the given video. After watermarking, the video is encoded again [16]. The extracted watermarks are compared with the features derived from the received video, to check the authenticity of the given video.

#### 4.1.3. Intelligent Techniques

Intelligent techniques for video authentication use database of videos. The database comprises authentic video clips as well as tampered video clips. As in [33], the authors proposed an intelligent technique for video authentication which uses inherent video information for authentication, thus making it useful for real world applications. The proposed algorithm in [33] is validated using a database of 795 tampered and non tampered videos and the results of algorithm show a classification accuracy of 99.92%. The main advantage of intelligent techniques is that they do not require the computation and storage of secret key or embedding of watermark. The algorithm in [33] computes the local relative correlation information and classifies the video as tampered or non-tampered. Here the algorithm uses Support Vector Machine (*SVM*) for the classification of the tampered and authentic videos. *SVM* [10] is a powerful methodology for solving problems in non linear classification, function estimation and density estimation [11]. This algorithm [33] is performed in two stages: (1) *SVM* training and (2) Tamper detection and classification, using *SVM*. In *SVM* training, the algorithm trains the *SVM* by using a manually labelled training video database, if the video in the training data is tampered, then it is assigned the label -1 otherwise the label is +1 (for the authentic video). From the training videos, relative correlation information between two adjacent frames of the video is computed, with the help of corner detection algorithm [14]. Then relative correlation information *RC* is computed for all adjacent frames of the video with the help of

$$RC = \frac{1}{m} \sum_{i=1}^m L_i$$

Where  $L_i$  is local correlation between two frames for  $i = 1, 2, \dots, m$ . and  $m$  is the number of corresponding corner points in the two frames. The local correlation information *RC* is computed for each video and the *RC* with the label information of all the training video data are provided as input to the *SVM*. With this information of all the video in video database, the *SVM* [10] is trained to classify the tampered and non tampered video data. Output of *SVM* training is a trained hyper plane with classified tampered and non tampered video data. In [24], authors integrate the learning based support vector machine classification (for tampered and non tampered video) with singular value decomposition watermarking. This algorithm is independent of the choice of watermark and does not require any key to store. This intelligent authentication technique embeds the inherent video information in frames using *SVD* watermarking and uses it for classification by

projecting them into a non linear *SVM* hyper plane. This technique can detect multiple tampering attacks.

#### 4.1.4. Other Authentication Techniques

Apart from digital signature, watermarking and intelligent techniques, various other techniques are proposed by researchers for authentication purpose of digital video in the literature. In [19], an authentication scheme for digital video is introduced which is based on motion trajectory and cryptographic secret sharing [19]. In this scheme, the given video is first segmented into shots then all the frames of the video shots are mapped to a trajectory in the feature space by which the key frames of the video shot are computed. Once the key frames are evaluated, a secret frame is computed from the key frames information of the video shot. These secret frames are used to construct a hierarchical structure and after that final master key is obtained. This master key is used to identify the authenticity of the video. Any modification in a shot or in the important content of a shot will be reflected as changes in the computed master key. Here trajectory is constructed, using the histogram energy of the frames of the video shot. Once the key frames are computed these are utilized to compute the secret frame by extrapolation. Now an interpolating polynomial  $f(x)$  is computed by using key frames as follows.

$$\sum_{j=1}^{n+1} \prod_{i=1}^{n+1} \frac{x-x_i}{x_j-x_i} I_j$$

This is Lagrange interpolation formulation where the  $x_i$  position refers to each key frame and  $I_i$  is the pixel value of the key frames. By using this equation and extrapolation a frame at  $x = 0$  is computed, which is regarded as the secret key. Considering the set of secret keys as another set of shares, the master key frame is computed for that particular video. With this scheme any video can be authenticated by comparing its computed master key with the original master key. This comparison can be performed by using the general cosine correlation measure given by:

$$sim = \frac{I_O \cdot I_N}{|I_O \cdot I_N|}$$

Where  $I_O$  and  $I_N$  are the original master key and the new master key considered as vectors. The similarity value would be in the range [0, 1] and if  $sim = 1$ , the two master keys would be the same, however if  $sim = 0$ , the two master keys would be different. In [20], the key frames are selected by deleting the most predictable frame. In the approach of [21], the key frames are extracted from a video shot based on the nearest feature line. The work in [22] authenticates a video by guaranteeing the edited video to be the subsequence of the original video using a special hash function. The MPEG video standard is one of the most popular video standards in today's digital era. In [35] Weihong Wang and Hany Farid have been worked on MPEG video standard (MPEG-1 and MPEG-2) in this paper they specifically show how a doubly compressed MPEG video sequence introduces specific static and temporal statistical perturbations whose presence can be used as evidence of tampering. In [36] Hany Farid describes three techniques to expose digital forgeries in



which the approach is to first understand how a specific form of tampering disturbs certain statistical properties of an image and then to develop a mathematical algorithm to detect this perturbation. These are Cloning, Lighting and Retouching. In Cloning, a digital image is first partitioned into small blocks of the regions. The blocks are then reordered so that they are placed a distance to each other that is proportional to the differences in their pixel colours [36]. Since it is statistically unlikely to find identical and spatially coherent regions in an image, therefore their presence can be used as evidence of tampering. In lighting approach the direction of an illuminating light source for each object or person in an image is automatically evaluated by some mathematical techniques. The retouching technique exploits the technology by which a digital camera sensor records an image, for detecting a specific form of tampering.

## 5. Challenging Scenarios for Video Authentication

In some of the surveillance systems storage and transmission costs are the important issues. In order to reduce the storage and transmission cost only those video clips which contain objects of interest are required to be sent and stored. Moreover in most of the surveillance applications, background object changes very slowly in comparison to foreground objects. A possible efficient solution in these scenarios is that only the objects of interest (mostly foreground objects) are sent out frame by frame in real time while the background object is sent once in a long time interval. In such surveillance applications, it becomes very critical to protect the authenticity of the video: the authenticity against malicious alterations and the authenticity for the identity of the transmission source (i.e. identifying the video source). In event based surveillance systems, the video sequences are captured when there is any kind of change in the scene (existence of an event) which would be captured by the camera. If there is uniformity in the scene in such a way that there is not any change in the scene then the surveillance camera does not capture any video sequence. This kind of surveillance system is used in military system for border security purpose. Authenticity for this kind of video sequences is a challenging issue because there is no proper time sequence in video sequences which are captured by surveillance camera. These are the scenarios which pose considerable challenges to the researchers for authentication.

## 6. Summary

Fig. 9 presents a tree structure of the methodologies that can be used for video authentication. The four children node of the root node covers almost all the methodologies. The leaf nodes of the tree structure show the key points of their grandparent node methodologies. This tree structure shows how all the methodologies use different approaches for video authentication. However many work has been done in watermarking and digital signature methodologies, other techniques (including intelligent technique) also

produce better results for authentication purpose. There is no issue related with the size of authentication code in digital signature techniques, however, they provide better results regarding robustness, since the digital signature remains unchanged when there is a change in pixel values of the video frames. But if the location where digital signature is stored is compromised then it is easy to deceive the authentication system. On the other hand fragile watermarking algorithms perform better than algorithm based on conventional cryptography [32]. Fragile and semi fragile algorithm show good results for detecting and locating any malicious manipulations but often they are too fragile to resist incidental manipulations. Moreover embedding the watermark may change the content of video which is not permissible in court of law [33]. In addition of these techniques, intelligent techniques explore the new dimensions in video authentication. However learning based intelligent authentication algorithm does not require computation and storage of any key or embedding of secret information in the video data, it requires a large database of tampered and non tampered video to learn the algorithm so that it can classify whether the given video is authentic or not. These techniques are slower than some existing authentication techniques, since they use sufficient large database to learn the algorithm. In other techniques, most of the authentication techniques are established for specific attacks. For example motion trajectory based algorithm only detects the frame addition and deletion attacks (temporal attacks). Moreover compression and scaling operations also affect the performance of existing algorithms.

## 7. Conclusion

Video authentication is a very challenging problem and of high importance in several applications such as in forensic investigations of digital video for law enforcement agencies, video surveillance and presenting video evidence in court of law. However with growing development in video editing tools and wide availability of these powerful editing software video tampering attacks explores new dimensions in various fields. In future it is going to be a big menace for information security. By analysing the various video authentication techniques that were presented in this paper we can say that the authentication techniques are specific to the applications (surveillance, entertainment industry, medical, copyright...). As the time passes, we are getting more involved with video applications, in our daily lives. Our information systems are greatly dependent on video applications, now. This, with a wide range of tampering attacks, causes severe challenges on information security. In future robustness would be the key point for video authentication techniques, so that it can differentiate the acceptable video processing operations from malicious tampering attacks. However A perfect video authentication algorithm that detects all kinds of malicious manipulations and that can tolerate all content preserving manipulations is yet to be discovered. We can hope for the better in the future.

## References:

- [1] B.G.Mobasseri, M.S.Sieffert, R.A.Simard, Content Authentication and tamper detection in digital video, Proc. IEEE International conference on Image Processing, Vancouver, September 10-13, 2000.
- [2] B.G.Mobasseri, A.E.Evans, Content dependent video authentication by self water marking in color space, Proc. Security and watermarking of multimedia contents III, vol. 4314 pp.35-46, January 21-26, 2001.
- [3] M.V. Celik et al, Video authentication with self recovery, Proc. Security and watermarking of multimedia contents IV vol. 4314, pp. 531-541, January 21-24, 2002.
- [4] Fabrizio Guerrini, Reccardo Leonardi and Pierangelo Migliorati A new video authentication template based on bubble random sampling.
- [5] Peng Yin, Hong heather Yu, Classification of Video Tampering Methods and Countermeasures using Digital Watermarking Proc. SPIE Vol. 4518, p. 239-246, Multimedia Systems and Applications IV
- [6] Pradeep K. Atrey, Wei-Qi Yan, Ee-Chien Chang, Mohan S. Kankanhalli, A hierarchical signature scheme for robust video authentication using secret sharing.
- [7] Johns Hopkins APL creates system to detect Digital Video Tampering. <http://www.jhu.edu/>
- [8] Ching-Yung Lin, Shih-Fu Chang, "Issues and Solutions for authenticating MPEG Video" SPIE electronic Imaging 1999. San Jose.
- [9] M. P. Queluz Authentication of digital images and video: Generic models and a new contribution.
- [10] Vapnik VN (1995) The nature of statistical learning theory. Springer Verlag.
- [11] Singh R., Vatsa M., Noore A (2006) Intelligent biometric information fusion using support vector machine. In soft computing in Image processing: Recent advances, Springer Verlag 327-350.
- [12] R. Gennaro and P. Rohatgi, How to sign digital stream, Crypto' 97, pp. 180-197, 1997.
- [13] J. M. Park, E. K. P. Chong and H. J. Siegel, Efficient multicast packet authentication using signature amortization, IEEE symposium on security and privacy, pp. 227-240, 2002.
- [14] Kovesei PD (1999) Image features from phase congruency. Videre: Journal of Computer vision research, MIT Press 1(3).
- [15] Navajit Saikia, Prabin K Bora, Video Authentication using temporal wavelet transform.
- [16] Chang-yin Liang, Ang Li, Xia-mu Niu Video authentication and tamper detection based on cloud model.
- [17] Ditmann, J.; Steinmetz, A; Steinmetz, R., Content based digital signature for motion pictures authentication and content fragile watermarking, Multimedia computing and systems, 1999. IEEE International Conference on, Volume: 2, 1999, Page(s): 209-213 vol. 2.
- [18] Queluz, M. P., Toward robust, content based techniques for image authentication, Multimedia signal processing, 1998 IEEE Second workshop on, 1998 page(s): 297-302.
- [19] Wei-Qi Yan an Mohan S Kankanhalli, Motion Trajectory Based Video Authentication *ISCAS (3) 2003*: 810-813
- [20] Latechi L. Wildt D. and Hu J., Extraction of key frames from videos by optimal color composition matching and polygon simplification. Proceedings of MMSP' 2000, Cannes, France, October 2001
- [21] Zhao L., Qi W., Li S., Yang S. and Zhang H., Key frame extraction and shot retrieval using Nearest Feature Line (NFL)., Proceedings of ACM Multimedia 2000.
- [22] Quisquater J., Authentication of sequences with the SL2 Hash function application to video sequences, Journal of computer security, 5(3), pp: 213-223, 1997.
- [23] Chun-Shien Lu and Hong Yuan Mark Liao, Structural digital signature for image authentication: An Incidental Distortion Resistant Scheme. *IEEE Trans. Multimedia*, vol. 5, no. 2, pp. 161-173, Jun. 2003.
- [24] R. Singh, M. Vatsa, S.K. Singh, and S. Upadhyay, Integrating SVM Classification with SVD Watermarking for Intelligent Video Authentication, In Telecommunication Systems Journal - Special Issue on Computational Intelligence in Multimedia Computing, Springer, 2008.
- [25] S. Bhattacharjee and M. Kutter, Compression tolerant image authentication, in IEEE International Conference on Image Processing, 1998, pp. 435-439.
- [26] W. Diffie and M. E. Hellman, New Directions in cryptography, IEEE Trans. on Information Theory, Vol. 22, No. 6, pp.644-654, Nov 1976.
- [27] P. Wohlmacher, Requirements and Mechanism of IT-Security Including Aspects of Multimedia Security, Multimedia and Security Workshop at ACM Multimedia 98, Bristol, U. K., Sep. 1998.
- [28] Shui-Hua Han, Chao-Hsien Chu, Content based image authentication: current status, issues, and challenges. Int. J. Inf. Security (2010) 9:19-32, DOI 10.1007/s 10207-009-0093-2.
- [29] S. Craver, N. Memon, B. Yeo and N. M. Yeung, Resolving Rightful Ownerships with Invisible watermarking Techniques: Limitations, Attacks and Implications, IEEE Journal on Selected Areas in Communications, Vol. 16, No. 4, pp. 573-586(1998).
- [30] The Oxford English Dictionary, 2<sup>nd</sup> Edition, Oxford University, pp. 795-796, 1989.
- [31] The Webster's New 20<sup>th</sup> Century Dictionary.
- [32] Adil Hauzia, Rita Noumeir (2007) Methods for image authentication: a survey. In: Proceedings of the Multimedia Tools Appl (2008) 39:1-46, DOI 10.1007/s11042-007-0154-3
- [33] S. Upadhyay, S.K. Singh, M. Vatsa, and R. Singh, Video authentication using relative correlation information and SVM, In Computational Intelligence in Multimedia Processing: Recent Advances (Springer Verlag) Edited by A.E. Hassanien, J. Kacprzyk, and A. Abraham, 2007
- [34] Jana Dittman, Anirban Mukharjee and Martin Steinbach Media independent watermarking classification and the need for combining digital video and audio watermarking for media authentication. International conference on Information Technology: Coding and Computing, 2000.
- [35] Weihong Wang, Hany Farid Exposing digital forgeries in video by detecting double MPEG compression.
- [36] Hany Farid, Digital doctoring: How to tell the real from fake
- [37] Po-chyi Su, Chun-chieh Chen and Hong Min Chang, Towards effective content authentication for digital videos by employing feature extraction and quantization
- [38] Pradeep K. Atrey, Abdulmotaleb El Saddik, Mohan Kankanhalli, Digital Video Authentication, IGI Global, 2009.

**Saurabh Upadhyay** received the B. Tech. degree in computer science and engineering in 2001 and is currently working toward the Ph.D. degree in computer science at U.P. Technical University, India. He is an Associate Professor in the Department of Computer Science and Engineering, Saffrony Institute of Technology Gujarat, India. He is actively involved in the development of a robust video authentication system which can identify tampering to determine the authenticity of the video. His current areas of interest include pattern recognition, video and image processing, watermarking, and artificial intelligence

**Sanjay K. Singh** is Associate Professor in Department of Computer Engineering at Institute of Technology, BHU, India. He is a certified Novel Engineer and Novel administrator. His research has been funded by UGC and AICTE. He has over 50 publications in refereed journals, book chapters, and conferences. His research interests include computational intelligence, biometrics, video authentication and machine learning. Dr. Singh is a member of IEEE, ISTE and CSI.

# Novel Design of A Compact Proximity Coupled Fed Antenna

Mehdi ALI<sup>1</sup>, Abdennacer KACHOURI<sup>2</sup> and Mounir SAMET<sup>3</sup>

<sup>1</sup>University of Sfax, National School of Engineers of Sfax,  
Laboratory LETI, Route Soukra Km 3.5 B.P W 1173, TUNISIA

<sup>2</sup>University of Gabes, ISSIG Higher Institute Of Industrial Systems Gabes  
Gabes CP 6011 TUNISIA

<sup>3</sup>University of Sfax, National School of Engineers of Sfax,  
Laboratory LETI, Route Soukra Km 3.5 B.P W 1173, TUNISIA

## Abstract

Certain applications such as RFID, on body sensors network, microwave systems usually require good matching impedance, high gain and large bandwidth for their antennas. The aperture coupled antenna is one candidate that can provide high gain large bandwidth and little packaging. Thus, it would be of interest to enhance the characteristics of a singly-fed aperture antenna used for Zigbee application.

In this paper, we are presenting a new design of aperture coupled rectangular patch antenna operating at 2.45 GHz ISM-band frequency. The objective of this design is not limited to the improvement of the impedance bandwidth but also to better the coupling involved. The cross-polarisation (X-pol), the backward radiation and the half-power beam width in two orthogonal planes are also examined.

The proposed design is based on a new aperture coupling technique in which two slots are fed by a microstrip line and coupled to a parasitic patch radiator etched on the opposite side from the slots. The matching impedance for a conventional aperture coupled microstrip antenna is obtained by the adjustment of the dimension of the slot and the feeding line. Here, the distance separating the slots is employed to control the coupling and modifying the input impedance of the antenna. Therefore, accurate matching impedance is reached with a good radiation pattern.

**Keywords:** *aperture coupled antenna, impedance matching, microstrip, coupling, multilayer, input impedance.*

## 1. Introduction

Conventional microstrip aperture antennas have at least two layers; on the higher layer is printed a conducting patch while the second is a grounded microwave substrate on which is etched the feeding line. In the other side, a slot opening is operated through which electromagnetic waves are transmitted or received.

The impedance bandwidth is a function of resonant frequency or patch size. On the other hand, the gain of

microstrip patch antennas has been shown to be a strong function of the substrate permittivity and thickness. Therefore, the enhancement of gain and bandwidths for a patch antenna are always challenging tasks and especially for aperture antenna that depend on the coupling generated by the aperture in the middle placed ground plane between the patch and the feeding line [1].

Much research effort has been made to investigate modified methods for decreasing the whole area and maintaining the initial performance. Previous methods have focused on increasing the electrical field on the open end for the uniform field on the aperture [2]. Hence, the associated studies with the H-shaped slot coupled antenna, dumbbell shaped slot coupled antenna, and bow-tie shaped slot coupled antenna [3] are adopted to improve the amount of electromagnetic coupling. The impedance matching is done by acting on the aperture dimensions and the stub length that also make a shifting in the resonant frequency [4], the electrical dimension of the patch should be corrected and a difficult conciliation is hard to achieve.

This paper describes a new aperture coupled antenna with two separate slots in order to reduce the area of antenna and achieve better performance. It also exhibits excellent efficiency and soft means to attain excellent impedance matching. A comparative study is done between a basic design and the new proposed configuration.

A three-layer antenna of an ordinary FR4 substrate is designed. The matching impedance is obtained by adjusting the gap between the two apertures and a reflexion coefficient S11 about 50dB is easily achieved for a reduction of about 10% of the total surface, an enhancement of the bandwidth of 20% and the gain rises by about 3.4%.

## 2. Aperture coupled antenna structure

Fig.1 shows a basic antenna structure with a rectangular patch which is excited through one slot on the ground plane. Typically the lower layer is a two metallised FR4, the feeding line should be exactly etched below the aperture a shifting in the line position compared to the aperture leads to a shifting in the resonant frequency and influences the coupling[5]. We assume the following structure to avoid fabricating and assembling constraints since it is difficult to print the feeding line exactly beneath the aperture centre. The antenna will be composed of three layers made of the same material with the same thickness; a one side metallised 1,6 mm thick FR4 substrate will be used. The lower layer has on the 50Ω feeding line, the middle layer is a grounded plane with the slot aperture and finally the higher layer on which is printed the radiating patch. The aperture, the feeding line and the patch centering is obtained by operating centring holes in the three layers. The structure is assembled using epoxy glue; a compact and resistant antenna is obtained.

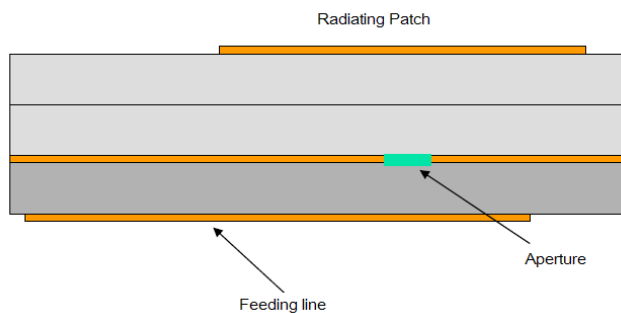


Fig.1 conventional aperture-coupled antenna

### 2.1 Conventional antenna testing.

An aperture coupled antenna could be replaced by the equivalent circuit shown in Fig.2. The resonant patch dimensions are determined using microstrip antenna theory equations. The slot introduces a capacitance  $C_s$  determined by considering the patch dimension, layers thickness, aperture dimension and feeding line length and width.

The aperture coupling consists of two substrates separated by a ground plane. On the bottom side of the lower substrate, there is a microstrip feed line which energy is coupled to the patch through a slot on the ground plane separating the two substrates. This arrangement allows independent optimization of the feed mechanism and the radiating element. Typically matching impedance is performed by controlling the length of the feed line and the length of the slot. The coupling through the slot can be modeled using the theory of Bethe [6],

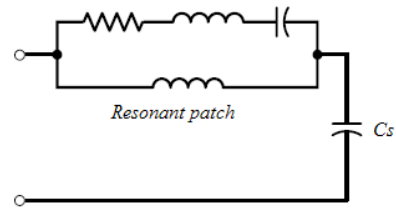


Fig. 2 The simplest equivalent circuit of an aperture coupled antenna

The 3D structure of a conventional aperture coupled antenna tested is showed on Fig.3.

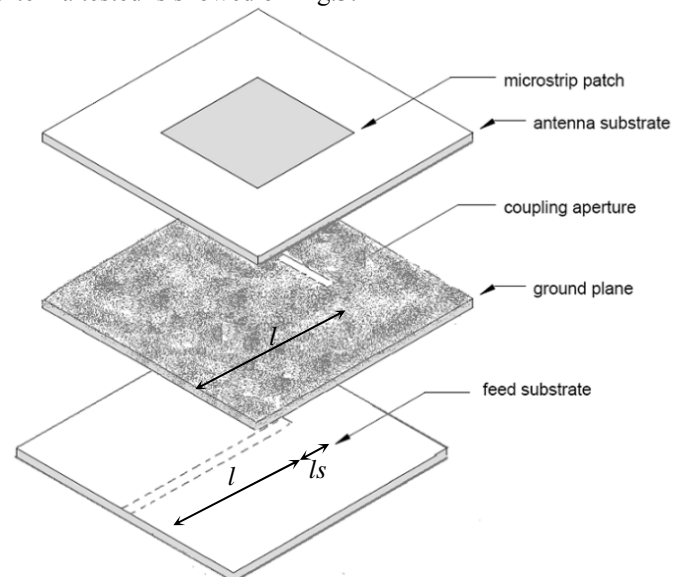


Fig.3 conventional aperture coupled antenna structure

The optimal distance  $l$  between the aperture and the feeding point is experimentally tested and determined as  $\lambda_g/2$ .

The mask of the conceived antenna is showed in Fig.4

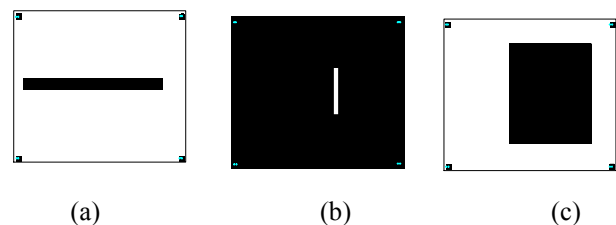


Fig. 4 (a) lower layer, (b) Middle layer and (c) higher layer  
 (The substrate FR4 1,6mm one side copper metallased 30μm  $\epsilon=4.4$ ,  $\tan\delta=0.018$ , total size (50x55)mm<sup>2</sup>)

The equivalent lumped element circuit is shown below.

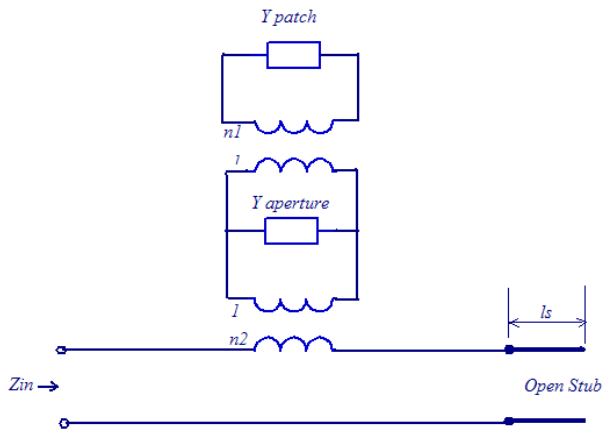


Fig.5 Lumped element equivalent circuit.

The above Figure highlights the coupling generated by the aperture. Many investigations have aimed to make the coupling better by implementing other aperture shapes. The Slot length affects the coupling level and the back radiation level [7]. For maximum coupling, the patch and the feeding line should be centered over the slot. Hence, it is desirable to use a shape that has maximum coupling for a given size. The feed line affects the level of the radiation by carrying energy from a port to the actual antenna and so to launch guided waves only.

A simulation of the conventional antenna is done using the ADS Advanced Design System. The antenna dimension and result are summarised in Tab.1 and Tab.2.

Table 1 Conventional Aperture Antenna dimension

<b>Patch</b>	W	31.5mm
	L	25.47mm
<b>Substrate</b>	Cr	4.4
	h	3x1.59mm
	Tanδ	0.018
<b>metallization</b>	t	35μm
	Metal Permeability	1
	Metal Conductance	1,83e+7
<b>Aperture</b>	Width ( $w_a$ )	1.35mm
	aperture large ( $l_a$ )	16.13mm
<b>Feed line</b>	Width ( $w_f$ )	3.41mm
	length ( $l_f$ )	25.84mm
<b>Stub</b>	Width ( $w_s$ )	3.41mm
	Length ( $l_s$ )	18.65mm

Table 2 Aperture conventional antenna performances

<b>Power radiated(watts)</b>	0.13
<b>Effective angle(degrees)</b>	162
<b>Directivity(dB)</b>	6.46
<b>Gain(dB)</b>	3.81
<b>Max. Intensity(w/Steradian)</b>	0.04
<b>Bandwidth (GHz)</b>	2.407-2.493

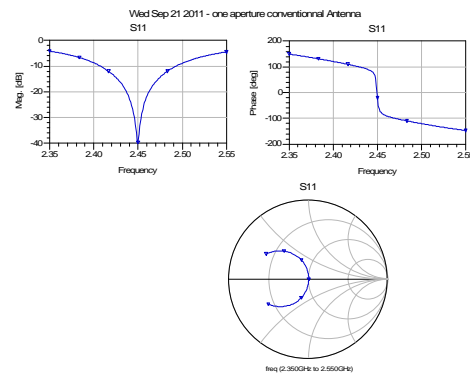


Fig.6 Simulation of the conventional aperture-coupled antenna

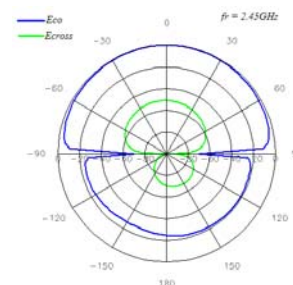


Fig. 7 Far fiel cut of The electric field Eco and Ecross of the conventional aperture coupled antenna

## 2.2 Enhanced gain and bandwidth aperture-coupled antenna.

The proposed enhancement of the coupling is supplied by adding another slot in the ground plane. The coupling is controlled by varying the distance between the slots. The antenna designed is illustrated in Fig 8

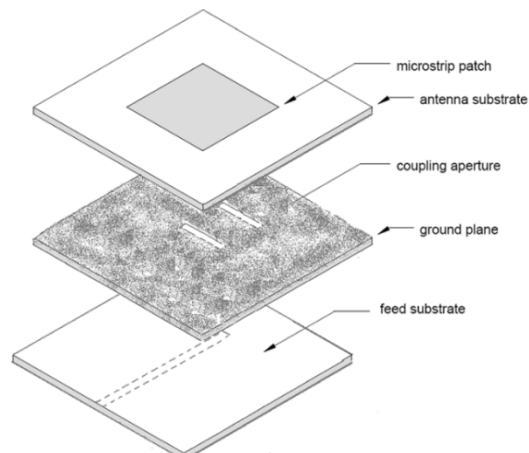


Fig. 8 3D new conceived apertures-coupling antenna.

The equivalent lumped element circuit shows that the effect of the supplementary aperture is to increase the inductive effect. The length of the stub needed to match the impedance will be shorter than the stub used for conventional antenna and so a reduction of the antenna dimensions will be obtained.

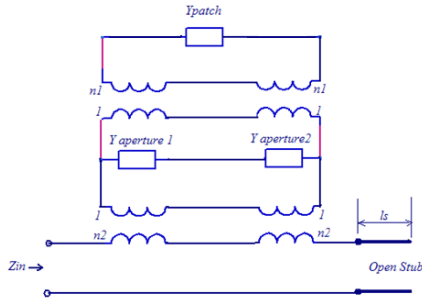


Fig. 9 Lumped element equivalent circuit two slots.

The mask of the designed antenna is shown in Fig.10.



Fig. 10 Masque of the conceived antenna

The simulation results are plotted below.

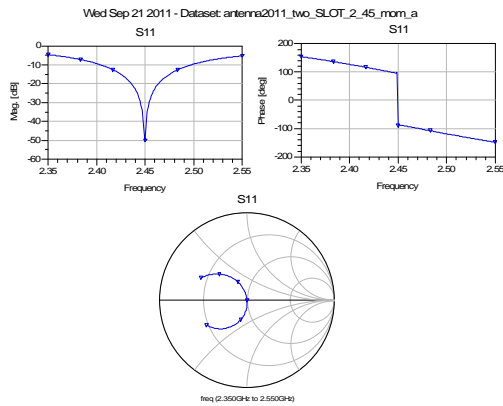


Fig. 11 Simulation results of the new conceived antenna.

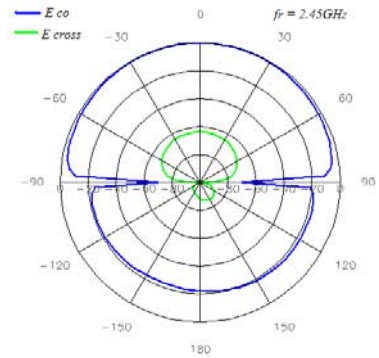


Fig. 12 Far field cut of The electric field Eco and Ecross

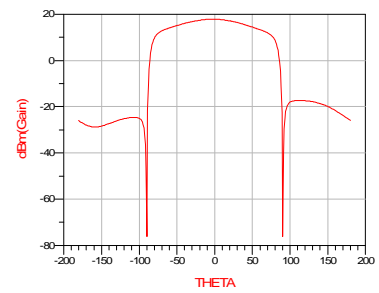


Fig. 13 The gain variation for the central frequency 2.45GHz

The antenna dimensions are placed in table 3.

Table 3 Double Apertures Antenna dimension

<b>Patch</b>	W	31.5mm
	L	26mm
<b>Substrate</b>	Er	4.4
	h	3x1.59mm
	Tanδ	0.018
<b>metallization</b>	t	35µm
	Metal Permeability	1
<b>Aperture</b>	Metal Conductance	1,83e+7
	Width ( $w_{ap}$ )	5.5mm
<b>Feed line</b>	aperture large ( $l_{ap}$ )	13.72mm
	Width ( $w_f$ )	3.41mm
<b>Stub</b>	length ( $l_f$ )	25.84mm
	Width ( $w_s$ )	3.41mm
	Length ( $l_s$ )	16.7mm

The radiating results obtained from a 3D Momentum simulation are summarised in Table.4.

Table 4 Double Apertures Antenna performance

<b>Power radiated(watts)</b>	0.13
<b>Effective angle(degrees)</b>	162
<b>Directivity(dB)</b>	6.46
<b>Gain(dB)</b>	3.94
<b>Max. Intensity(w/Steradian)</b>	0.04
<b>Bandwidth (GHz)</b>	2.4 - 2.5

Both conventional antenna and our design are fabricated a photo shown that a reduction of about 20% of the total surface was obtained with a notable enhancement of the antenna radiation characteristics.



Fig. 14 On the left the conventional design on the right our design

The measurement of the reflection coefficient of the double aperture antenna is given in Fig.15

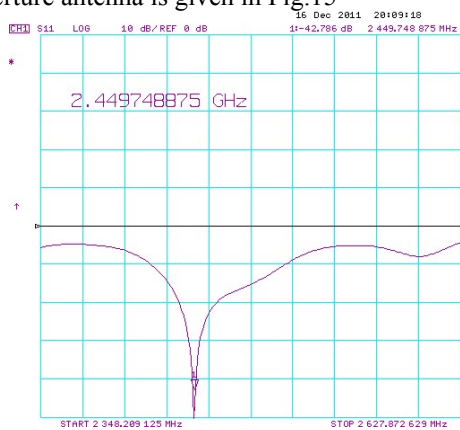


Fig. 15 S11 of the double apertures antenna

### 2.4 Comparative study:

We are comparing the finest tuning of the central frequency and the impedance matching of the conventional aperture coupled antenna and our design.

For both designs the inductive effect is generated by controlling the stub length as shown in the simulation done for different stub length.

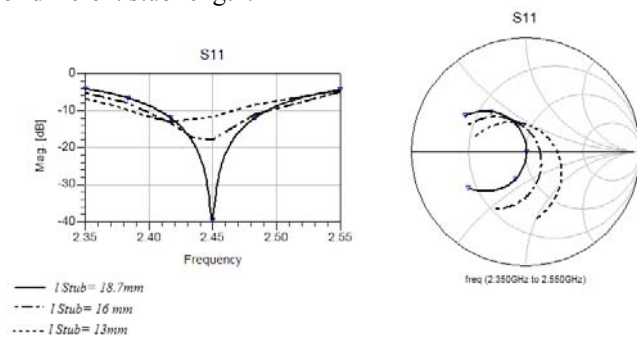


Fig. 16 effect of the stub tuning

The capacitive effect generated by controlling the aperture length for the conventional design is very sensible. The tuning of the length also affects the central resonant frequency very much as demonstrated in Fig.16 which summarizes the simulation for different aperture lengths.

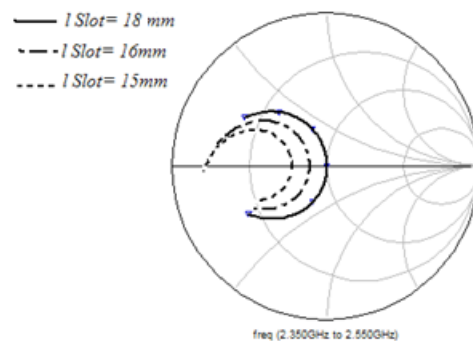
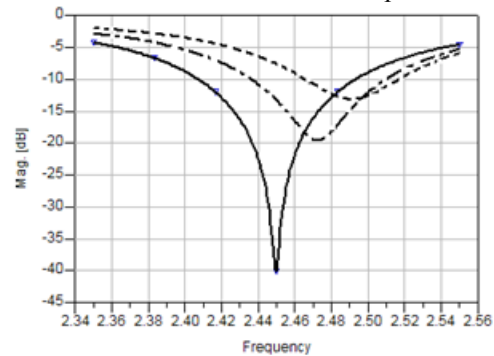


Fig. 17 Variation of the input impedance for differents aperture length.

For a conventional antenna, as the stub increases in length, the input impedance at a fixed frequency approximately follows a constant-resistance circle in Fig., with the reactance increasing according to the reactance of the open-circuited stub. The effect of increasing the aperture size is similar to that of increasing the size of the coupling power from a waveguide to a resonant cavity. When the aperture is small, the patch is under-coupled and the resonant resistance is less than the characteristic impedance of the feed line. As the aperture size increases, the coupling and the resonant resistance increase.

A wide range of resistance and reactance values can be achieved by adjusting the aperture length and the stub length.

The Fig. 17 shows also that the resonant frequency is very sensible to the aperture length and this, on the one hand makes the antenna impedance matching and central frequency adjustment very difficult and needs many trials. On the other hand, an error due to the fabricating process leads to important impedance miss-matching.

The same work is done for the new design. The tuning of the input impedance here is controlled by adjusting the distance separating the two slots and we conclude from the result obtained in Fig.17 that the central frequency is 50% slightly shifted compared to the conventional design.

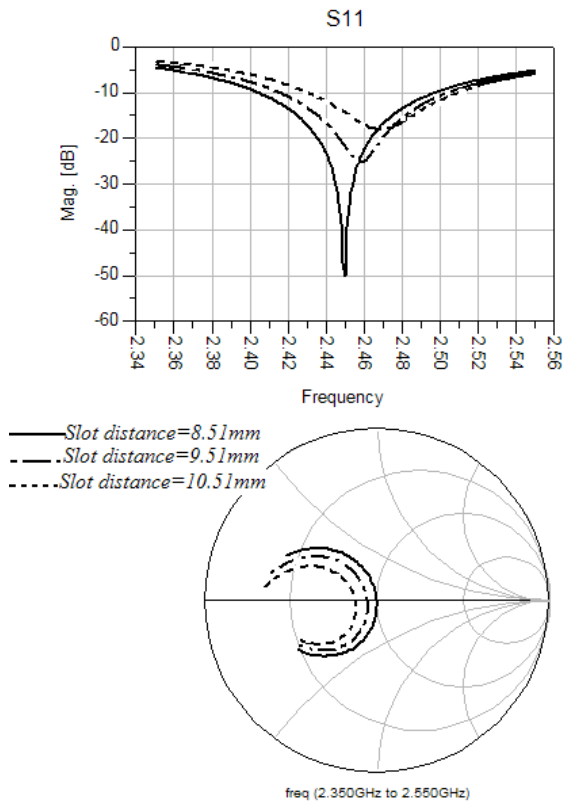


Fig.18 Variation of the input impedance and the centrale resonant frequency for different distance between the two slots.

The method proposed here is to operate two slots separated with a distance 'd', the slots consent to better the coupling leading to a better Gain and the variation of the distance 'd' controls the input impedance. So, to accurately match the input impedance of the antenna to that of the feeding source, it is enough to calculate the right length of the stub and the slot and to tune the coupling between the two slots by controlling the distance between them as illustrated in Fig.18.

### 3. Dimension determination

In this section we will detail the theoretical study of dimension determination of the antenna elements.

#### 3.1 The slots dimension

The slot length  $l_{ap}$  is chosen large enough so that sufficient coupling exists between the patch and the feed line but small enough so that it does not resonate within the band of operation, which usually leads to a significant radiated back lobe; a close formulation is set by [8].

$$l_{ap} = \frac{0.4\lambda_{eff}}{\sqrt{\frac{\epsilon_{eff(f)} + \epsilon_{eff(p)}}{2}}} \quad (1)$$

Were:

$\epsilon_{eff(f)}$  is the effective relative permittivity of the feed layer.  
 $\epsilon_{eff(p)}$  is the effective relative permittivity of the patch layer.

The slot width is usually chosen to be narrow to avoid a large back lobe component.

$$w_{ap} = 0.0164 \lambda_{eff} \quad (2)$$

#### 3.2 The feeding line dimension

The line characteristic impedance is taken  $50\Omega$  determined using Eq3.

$$Z_0 = \frac{87}{\sqrt{\epsilon_r + 1.4}} \ln \left( \frac{5.98h}{0.8w + t} \right) \quad (3)$$

This equation is valuable for  $h < 0.8w$ .

Where

$\epsilon_r$  = dielectric constant of the feeding layer.

$h$  = height of the substrate.

$w$  = width of the feeding line

$t$  = metallisation thickness

The feeding line is divided in two distinct parts, a half wave transformer and a matching stub. The total electric length is expressed as  $\varnothing_{eff} = 230^\circ$

The patch is centred at  $(x_0, y_0)$ , tested to be the optimal point giving the maximum coupling taken  $x_0 = L_{patch}$  and  $y_0 = 0$ .

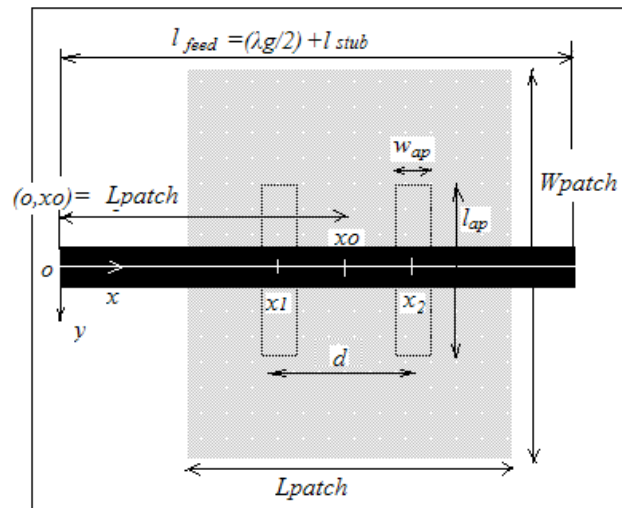


Fig. 19 position of antenna components

The stub is an open end; a tuning is necessary by controlling the stub length which can be modelled by an inductance [6] gives an accurate formula of L (nH).

$$L = 2 \cdot 10^{-4} \left[ \ln \left( \frac{ls}{w+t} \right) + 1,193 + 0,2235 \left( \frac{w+t}{ls} \right) \right] K_g \quad (4)$$

$$K_g = 0,57 - 0,145 \ln \frac{w}{h} \quad \text{for } \frac{w}{h} > 0,05 \quad (5)$$

For this study, we will consider the following fixed parameters:



$w_{ap}=2,67\text{mm}$ ,  $l_{ap}=13,7\text{mm}$ ,  $\epsilon_r(\text{FR4})=4,4$ ,  $\tan\delta(\text{FR4})=0,018$ ,  
 $\epsilon_r(\text{AR300})=2,2$ ,  $\tan\delta(\text{FR4})=0,0018$ ,  $\epsilon_r(\text{Air})=1$ ,  $\tan\delta(\text{Air})=0$ ,  
 $f_r=2,45\text{GHz}$ .

### 3.3 The patch dimension

The patch is a half wave resonator and is calculated using well known equations. The antenna patch dimension should be corrected. A reduction of 17% of the patch length should be applied due to the capacitive loading generated by the apertures.

The central frequency depends on many factors namely the patch dimension, the slot dimension, the slot position and the slot number.

### 3.4 The slots location

Conventionally, the aperture is located under the centre of the patch. This choice gives the maximum coupling but leads to less band width as demonstrated in the survey carried out here. For this design the slots are approximately located over the  $50\Omega$  impedance patch positions. The electric field collected by the two slots is superior to that collected by one slot located in the centre patch. Experiments are carried out to determine the distance separating the two slots for different aperture antennas made up of three layers. The first is made up of two layers of 1,6mm thick FR4 substrate the intermediary is an FR4 having variable thickness. The second, the higher and the lower layers are made up of 1,6mm thick FR4 substrate and the intermediary is variable thickness layer of air. The higher and the lower layers of the third antenna are made up of 1,6mm thick Arlon 300 substrate; the intermediary is variable thickness layer of air. Finally, we will test the effect of varying the wideness  $W$  of the rectangular patch on the distance  $d$ .

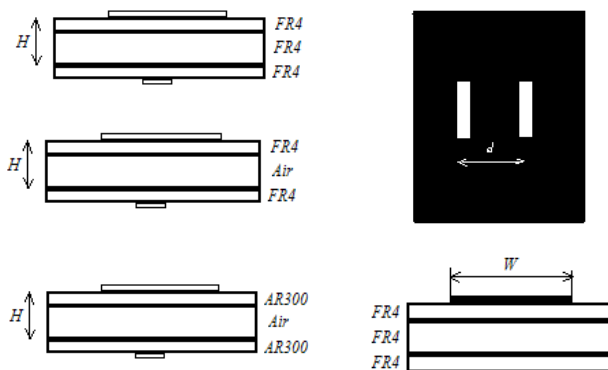


Fig. 20 Cases studied variation of  $H$  for different permittivity and variation of the wideness  $W$ .

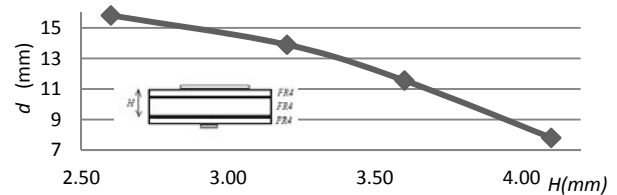


Fig.21 Aperture Antenna made up: FR4 - FR4 -FR4 ( $W=31,5$ ,  $L=26\text{mm}$ )

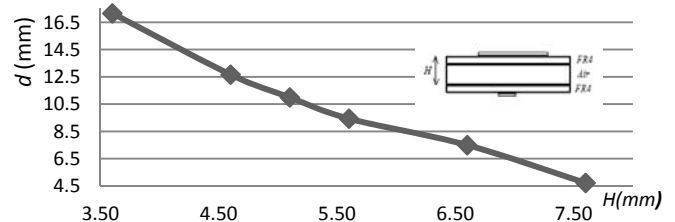


Fig. 22 A perture Antenna made up : FR4 - Air -FR4 ( $W=31,5$ ,  $L=26\text{mm}$ )

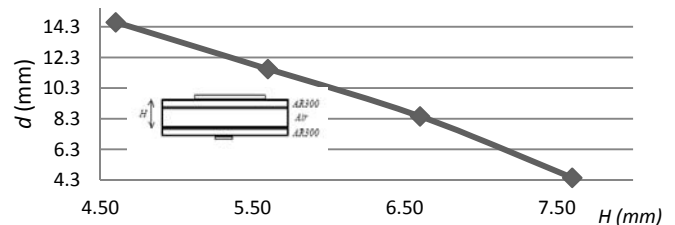


Fig.23 Aperture antrenna made up : Arlon300- AIR-Arlon300 ( $W=31,5$ ,  $L=26\text{mm}$ )

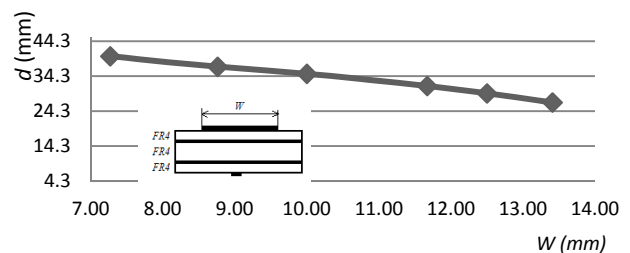


Fig.24 variation of  $d$  verses  $W$  for Aperture Antenna made up: FR4 - FR4 -FR4 ( $L=26\text{mm}$ )

From the experiments carried out, we can conclude that the distance 'd' separating the two slots depends on the coupling between the patch and the slots by varying the intermediary layer thickness, and depend on the input impedance of the patch by varying the permittivity and the patch wildness. An empirical formula giving the distance  $d$

separating the slots was determined and seems to produce accurate results for the cases studied.

$$d = \left( 1 - \left( 0,295 \log \left( \frac{\epsilon_r(\text{int}) + \epsilon_r(\text{p})}{2} \right) + \log \frac{H}{2,15} + \log \frac{0,67W}{l_{\text{ap}}} \right) \right) L \quad (6)$$

The parameters  $d, \epsilon_{\text{eff}(\text{p})}, \epsilon_{\text{eff}(\text{int})}, L$  are determined in Fig.23.

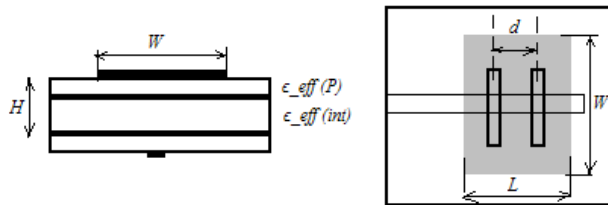


Fig. 23 Configuration of an aperture coupled antenna

Table . 5 slots distance separation for different thickness antenna made up of three layers FR4,FR4,FR4

H	2,6	3,2	3,6
<i>d simulated</i>	14,4	11,7	8,8
<i>d computed</i>	14	11,6	10

Table. 6 slots distance separation for different thickness antenna made up of three layers FR4,AIR,FR4

H	3,6	4,6	5,1	5,6	6,6	7,6
<i>d simulated</i>	17,8	13,7	11,0	9,4	7,5	4,7
<i>d computed</i>	17	13,1	11,4	10	7,4	5,1

Table. 7 slots distance separation for different thickness antenna made up of three layers Arlon300,Air,Arlon300

H	4,6	5,6	6,6	7,6
<i>d simulated</i>	14,6	11,5	8,4	6,9
<i>d computed</i>	17	13,4	10,6	8

Table. 8 slots distance separation for different wideness W antenna made up of three layers FR4-FR4-FR4

W	26,8	29,4	31,5	35,0	37,0	40,0
<i>d simulated</i>	13,41	12,5	11,67	10,0	8,76	7,27
<i>d computed</i>	13,5	12,5	11,76	10,6	9,15	8,7

From the tables we can conclude that a good agreement was found between the computed values of d and the simulated values.

## 4. Conclusions

The construction of a new generation of aperture coupled antennas is explained step by step with calculation details and then sensitivity analysis of the enhancement performed on the gain, bandwidth and size reduction. The

most important particularity is to use two slots to control the input impedance by tuning the coupling between them. The theoretical study presented could be used by designer to conceive such antenna; the separating gap between the slots giving the best impedance matching empirical formula presented here gives good agreement between the computed values and the simulated values.

A comparative study was carried out and confirmed enhancement; a reduction of about 10% of total surface, an improvement of the band width of 20% and the gain rises to about 3.4%.

## References

- [1] C. Hertleer, F. De Clercq, A. Tronquo, H. Rogier, and L. Van Langenhove, "Aperture-Coupled Patch Antenna for Integration Into Wearable Textile Systems," IEEE Antennas And Wireless Propagation Letters, Vol. 6, 2007
- [2] Kwok L. Chung' and Ananda S. Mohan, "Gain and Bandwidth Enhancement of a 2.4GHz Singly-Fed Cross-Aperture Coupled Patch Antenna," IEEE Xplore. pp. 410-413, 2002.
- [3] Qinjiang Rao, Tayeb A. Denidni, "A New Aperture Coupled Microstrip Slot Antenna" IEEE Transactions on Antennas And Propagation, Vol. 53, No. 9, September 2005.
- [4] Balanis, C.A. Antenna Theory Analysis and Design, Second Edition. United States of America. John Wiley & Sons 2005., p734.
- [5] P. Paul, J. S. Roy, S. K. Chowdhury Some experimental investigations on aperture-coupled microstrip antennas at TM<sub>11</sub> mode Microwave and Optical Technology Letters 22 MAR 2007
- [6] Ramesh Garg,Prakash Bhartia, Inder Bahl, Apisak Ittipiboon. Microstrip antenna design handbook 2001 ARTECH HOUSE.
- [7] A.ZARREEN , S.C.SHIVASTAVA " An Introduction of Aperture Coupled Microstrip Slot Antenna International Journal of Engineering Science and Technology Vol.2(1), 2010, 36-39.
- [8] Leung, K.W., et al., "Theory and Experiment of an Aperture-Coupled Hemispherical Dielectric Resonator Antenna," IEEE Transactions on Antennas & Propagation, Vol. 43, No. 1, Nov. 1995, pp. 192-198.

# ECG Analysis based on Wavelet Transform and Modulus Maxima

Mourad Talbi<sup>1</sup>, Akram Aouinet<sup>2</sup>, Riadh Baazaoui<sup>3</sup> and Adnane Cherif<sup>4</sup>

<sup>1</sup> High School of Applied Mathematics and Informatics of Kairouan, University of Kairouan  
Kairouan, Tunisia

<sup>2</sup> Faculty of Sciences of Tunis, University Tunis El-Manar  
Tunis, 1060, Tunisia

<sup>3</sup> Faculty of Sciences of Tunis, University Tunis El-Manar  
Tunis, 1060, Tunisia

<sup>4</sup> Faculty of Sciences of Tunis, University Tunis El-Manar  
Tunis, 1060, Tunisia

## Abstract

In this paper, we have developed a new technique of P, Q, R, S and T Peaks detection using Wavelet Transform (WT) and Modulus maxima. One of the commonest problems in electrocardiogram (ECG) signal processing, is baseline wander removal suppression. Therefore we have removed the baseline wander in order to make easier the detection of the peaks P and T. Those peaks are detected after the QRS detection. The proposed method is based on the application of the discretized continuous wavelet transform (Mycwt) used for the Bionic wavelet transform, to the ECG signal in order to detect R-peaks in the first stage and in the second stage, the Q and S peaks are detected using the R-peaks localization. Finally the Modulus maxima are used in the undecimated wavelet transform (UDWT) domain in order to detect the others peaks (P, T). This detection is performed by using a varying-length window that is moving along the whole signal. For evaluating the proposed method, we have compared it to others techniques based on wavelets. In this evaluation, we have used many ECG signals taken from MIT-BIH database. The obtained results show that the proposed method outperforms a number of conventional techniques used for our evaluation.

**Keywords:** Baseline drift, Continuous Wavelet Transform, Electrocardiogram, Modulus maxima, Thresholds, Window analysis.

## 1. Introduction

The electrocardiogram is the electrical activity signal of the heart. This activity is measured and recorded for more than a hundred years. The ECG analysis has been widely used for many cardiac diseases diagnosing. The ECG is a graphic record of the magnitude and detection of the electrical activity that is

generated by depolarization and repolarization of the ventricles and atria. In an ECG signal, one cardiac cycle consists of the P-QRS-T waves. The majority of the clinically useful information in the ECG is found in the amplitudes and intervals defined by its features (characteristics wave peaks and time durations). The development of quick and accurate techniques for automatic ECG feature extraction is of major importance, principally for the analysis of long recordings (Holters and ambulatory systems) [1]. In effect, the detection of the beats is necessary for heart rate determination and several related arrhythmias such as Bradycardia, Tachycardia and Heart Rate Variation; it is also necessary for further signal processing in order to detect abnormal beats [2]. The ECG feature extraction system provides fundamental features (amplitudes and intervals) to be used in subsequent automatic analysis.

All methods used by scientists are to help cardiologists to gain time, to interpret results and to improve the diagnostic. Though, ECG signals are characteristically corrupted by noise from electric interference, baseline wandering, and electromyography [3]. Therefore processing is necessary to cancel these noises while conserving information. The baseline drift in ECG signals, which often comes from the loose contact between skin and electrodes, also originates in the movements and breathe activity of patients [4].

In this paper, we proposed a technique using modified discretized continuous wavelet Transform, MMycwt, hard thresholding to detect R peaks, baseline wander removal and noise are suppressed, a varying-length window that is moving along the whole signals and modulus maxima based wavelet analysis employing the undecimated wavelet transform in order to detect and measure P and T waves.

## 2. Materials

### 2.1 Wavelet Transform

The theory of the wavelet transform (WT) is based on signal processing and developed from the Fourier transform basis. The wavelet transform is expressed as a series of functions which are related with each other by translation and simple scaling. The original WT function is called mother wavelet [5, 6] and is employed for generating all basis functions. A set of functions is constructed by scaling and shifting the mother wavelet  $\psi(t)$ . Those functions are expressed as follow:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right) \quad (1)$$

where  $a \in \mathbb{R}_+ - \{0\}$ ,  $b \in \mathbb{R}$  and  $b \in \mathbb{R}$ .

The original signal can be reconstructed by an appropriate integration and this is performed after projecting the given signal on a continuous family of frequency bands. The continuous wavelet transform (CWT) of a signal  $x(t)$  is given by:

$$CWT(a, b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} x(t) \psi^*\left(\frac{t-b}{a}\right) dt \quad (2)$$

where the superscript  $*$  is the complex conjugate and  $\psi_{a,b}^*$  represents a translated and scaled complex conjugated mother wavelet.

The mother wavelet  $\psi$  is invertible when it verifies the condition of admissibility which is stated as:

$$\int_{-\infty}^{+\infty} \frac{|\hat{\psi}(\omega)|}{\omega} d\omega < \infty \quad (3)$$

Many mother wavelets are used for computing the wavelet transform and Morlet is one of them. It is expressed as follow [7]:

$$\psi(t) = \frac{1}{\sqrt{\pi f_b}} e^{2i\pi f_c \cdot t} e^{-\frac{t^2}{f_b}} \quad (4)$$

where  $f_b$  and  $f_c$  are respectively a bandwidth parameter and a wavelet center frequency.

### 2.1 Modulus maxima

Wavelet modulus maxima are used for location characterizing singularities in the signal.  $wf(x)$  is the wavelet transform of a function  $f(x)$ .

- Any point  $x_0$  such that  $\frac{d(wf(x))}{dx}$  has a zero crossing at  $x = x_0$  is called a local extremum; when  $x$  varies.

- Any point  $x_0$  such that  $|wf(x)| \leq |wf(x_0)|$  when  $x$  belongs to the other side of the neighbourhood of  $x_0$ , and  $|wf(x)| < |wf(x_0)|$  when  $x$  belongs to either a right or left neighborhood of  $x_0$  is called modulus maximum.

- Any corrected curve in the scale space  $x$  along which all points are modulus maxima is called maxima line [8].

### 2.2 Database

The data available from MIT-BIH Arrhythmia Database [9] is the standard used by many researchers. The MIT-BIH database contains many data sets of electrocardiogram signals, mostly abnormal or unhealthy electrocardiograms, but it also contains normal electrocardiograms that can be used as a reference base [10]. This contains two lead ECG signals of 48 patients. The selected Arrhythmias are Premature Atria Beat (PAB), Premature Ventricular Beat (PVB), Right Bundle Branch Block (RBBB), and Left Bundle Branch Block (LBBB).

## 3. The proposed detection method

In this section, we have developed and evaluated a new detection method of P, Q, R, S and T Peaks. All the steps of the proposed technique are given as follow:

**Step1:** We apply the bionic wavelet transform (BWT) to the input ECG signal.

**Step2:** We smooth the bionic wavelet coefficients: each bionic wavelet coefficient is smoothed by using recTI.

**Step3:** We apply the inverse of the BWT to the smoothed bionic wavelet coefficients in order to obtain the smoothed ECG signal.

**Step4:** We apply the modified continuous wavelet transform (MMycwt) to the smoothed ECG signal. In this work we have modified Mycwt according to characteristics of the ECG signal in order to obtain the MMycwt. Note that the Mycwt is the descritized continuous wavelet transform used for the Bionic Wavelet Transform (BWT).

**Step5:** We apply hard thresholding to the forth wavelet coefficient in order to detect R peaks.

**Step6:** We use R-peaks for Q and S detection and this is performed by using the method of Mahmoodabadi et al [1]. This technique consists in searching for minimum of the signal about the R-peak within 0.1 second and this for detecting the Q and S peaks.

**Step7:** We suppress baseline wander removal in order to make easier the detection of P and T peaks.

**Step8:** We use a varying-length window that is moving along the whole signals (start window =  $S_i$ , end window =  $Q_{i+1}$ ).

**Step9:** Reduce the length of each window by eliminating the  $S_i$  and  $Q_{i+1}$  waves in order to make easier the detection of  $T_i$  and  $P_{i+1}$  waves.

**Step10:** The  $T_i$  and  $P_{i+1}$  waves are small in amplitudes so we increase the amplitude of the waves  $T_i$  and  $P_{i+1}$  by multiplying only the positive samples of them by an appropriate factor. This is done for the purpose to use a multi-scale product in Undecimated Discret Wavelet Transform (UDWT) domain.

**Step11:** We Apply the Undecimated Discret Wavelet Transform (UDWT) to each modified varying-length window.

**Step12:** Compute a Multi-scale product.

**Step13:** We take the modulus maxima of the obtained multi-scale product in order to detect P and T waves.

Step1, Step2 and Step3 constitute a new proposed technique of ECG denoising. As previously mentioned, we apply the procedure recTI to each noisy bionic wavelet coefficient in order to smooth it. The noisy bionic wavelet coefficients are obtained from the application of the BWT to the noisy ECG signal. Smoothing using the Translation-Invariant procedure (recTI), consists in applying threshold on the Forward Wavelet Transform Translation Invariant (FWT\_TI) coefficients [11]. Fig. 1 summarizes the main steps of the smoothing procedure, recTI.

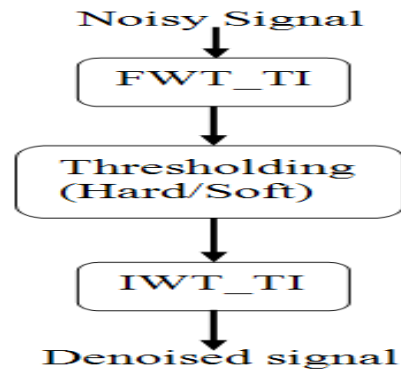
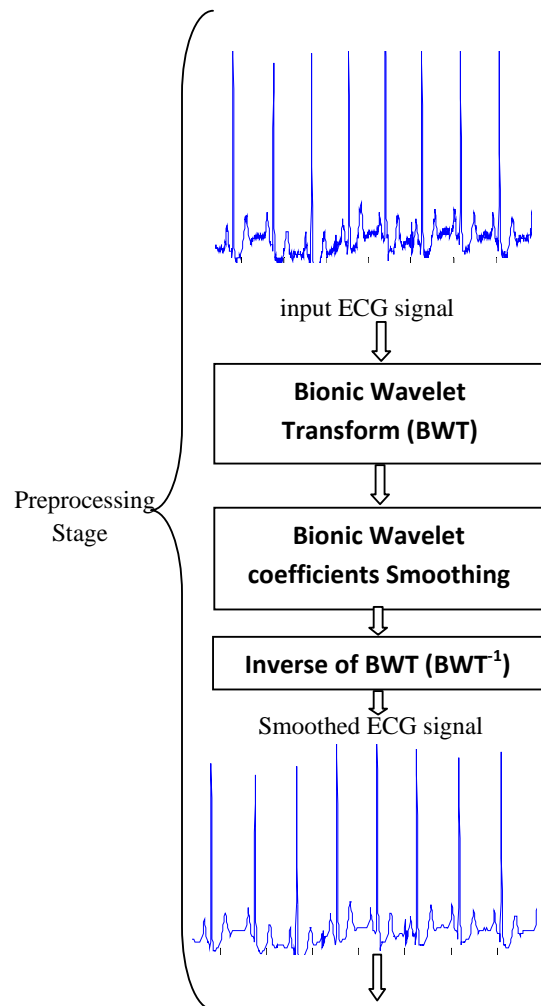


Fig. 1 Procedure smoothing by recTI.

All the previously mentioned steps of the proposed ECG peaks detection method are summarized in fig. 2.



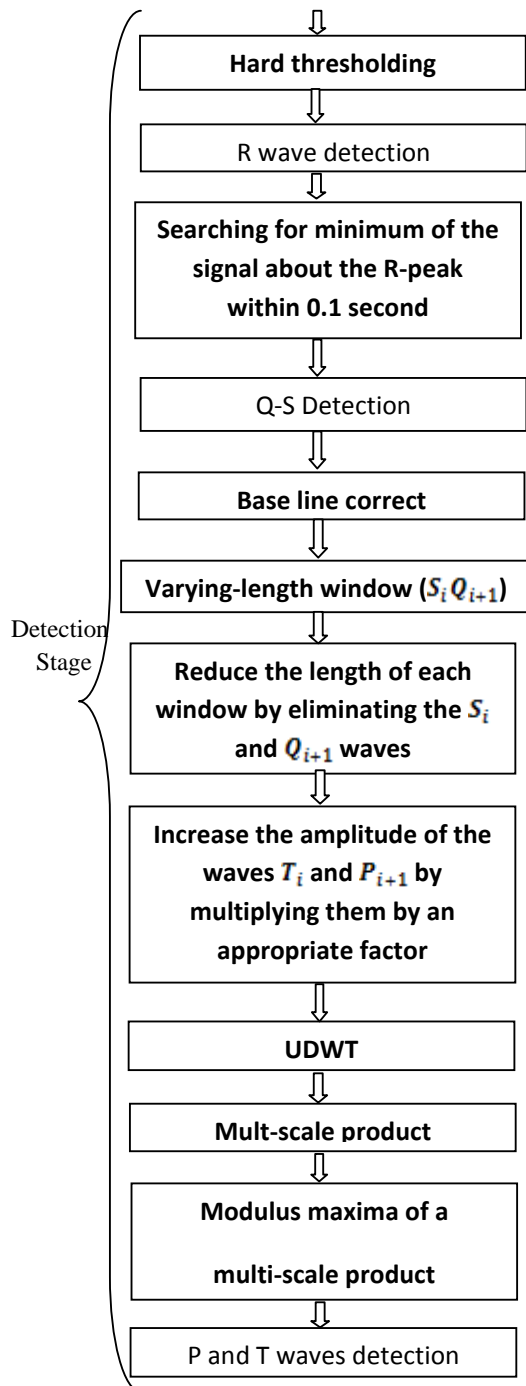


Fig. 2 A process of the proposed P, Q, R, S, and T wave detection.

### 3.1 Modified Mycwt (MMycwt)

For an ECG signal, the most important feature is the frequency range in which its main components occur [12]. Despite the existence of some other components like VLPs, we are interested in this paper in P, Q, R, S and T waves such as in the reference [12]. In references [13, 14], the value of  $f_0$  (the initial center frequency of the mother

wavelet) is equal to 15165.4Hz. As the scale increases, the center frequency goes smaller and smaller in the following way:

$$f_m = f_0/q^m, q > 1, m = 1, 2, \dots \quad (5)$$

We do not need such high frequency  $f_0$  for ECG signals. Omid et al [12] have optimized the value of  $f_0$  by running the program for different values of  $f_0$  and then minimizing the gradient of error variance by comparing the results numerically and morphologically with each other. It has been found that if  $f_0$  belongs to the range of 360 to 500Hz there would be no much distortion on the analyzed ECG signals [12]. In their work, Omid et al [15] have chosen 400Hz as the value of  $f_0$ . Hence, in our work, we have chosen  $f_0 = 250$  in order to obtain the MMyCwt. This choice of  $f_0$  yields satisfactory results. In this paper, we have chosen the value 1.1623 as that of  $q$  such as in the reference [13, 14].

### 3.2 Hard thresholding

After applying the MMyCwt to the input ECG signal, the fourth wavelet coefficient  $C4$  is thresholded using hard thresholding:

$$\text{if } (abs(C4(i)) \leq Thr) \text{ Then } C4(i) = 0$$

The threshold is selected to be:

$$Thr = \alpha \times \max(C4) \quad (6)$$

where  $\alpha$  is an appropriate positive parameter less than 1.

### 3.3 Baseline wander removal

One of the commonest problems in ECG signal processing is baseline wander removal and noise suppression, which determine posterior signal process. The amplitude of a wave is measured with reference to the ECG baseline level.

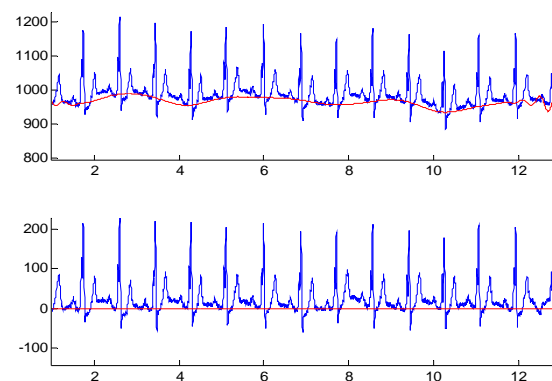


Fig. 3 Baseline corrected.

### 3.4 Varying-length window

After R-peaks, Q-peaks and S-peaks detection, we apply a varying-length window that is moving along the whole signals:

Window 1:  $[S_1, Q_2]$

Window 2:  $[S_2, Q_3]$

.....

.....

Window N:  $[S_N, Q_{N+1}]$

Each of those windows is modified by reducing its length and this is done by eliminating the  $S_i$  and  $Q_{i+1}$  waves. Then we have increased the amplitude of the waves  $T_i$  and  $P_{i+1}$  by multiplying only the positive samples of them by an appropriate factor.

### 3.5. Undecimated Wavelet Transform and modulus maxima

Finally, we have applied the Undecimated Discret Wavelet Transform (UDWT) to each modified varying-length window in order to compute the multi-scale product and then compute its modulus maxima as in [15] in order to detect the P and T peaks. The multi-scale product is calculated from the product of undecimated wavelet coefficients of successive scales (scale1, scale2 and scale3). The undecimated wavelet coefficients are obtained from Undecimated Discret Wavelet Transform (UDWT) application to each modified frame.

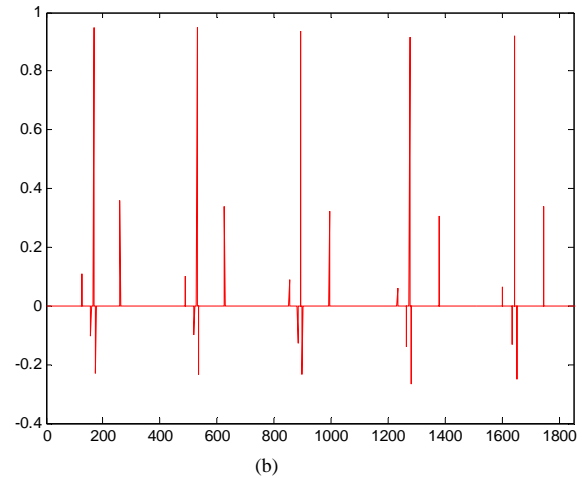
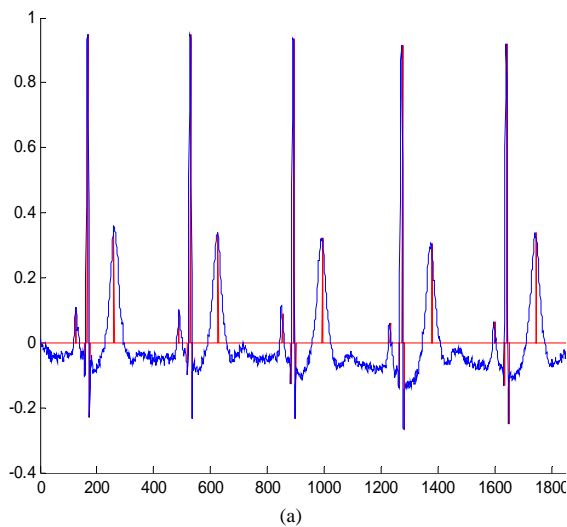
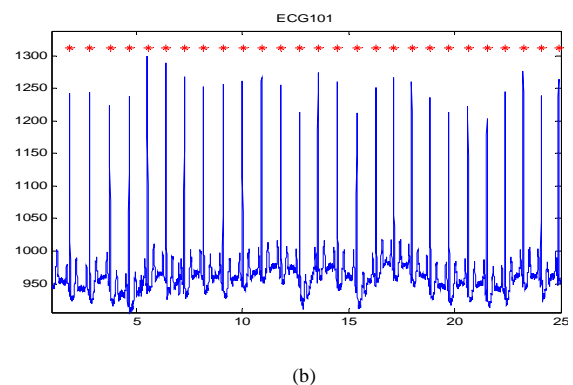
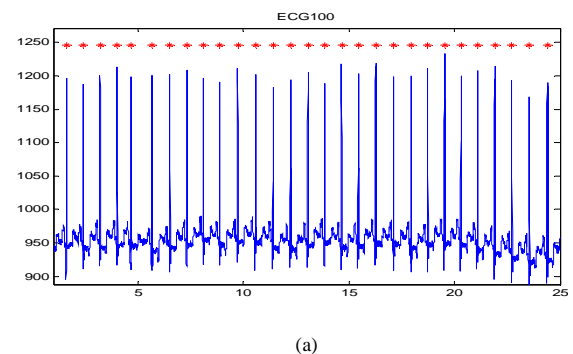


Fig. 4 (a) Original ECG signal (b) Positions of P, Q, R, S, and T peaks.

## 4. Results

The proposed algorithm has been validated on the MIT-BIH arrhythmia database to evaluate the P, QRS, and T detection. The database consists of 48 recordings; we use 46 half-hour recordings for a total 23 hours of ECG data. In first stage the positions of the R peaks have been detected and marked on the original signal. Fig. 5 shows some examples of ECG signal (color in blue) and the detection of R-peaks (in red color).



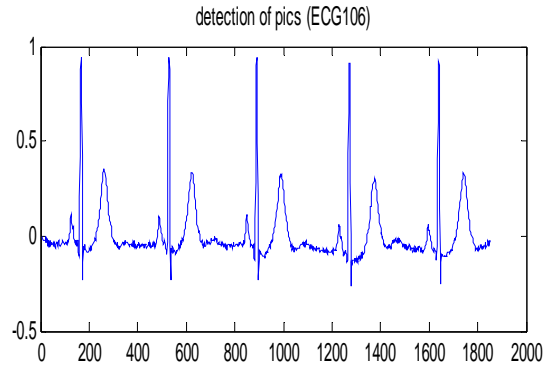
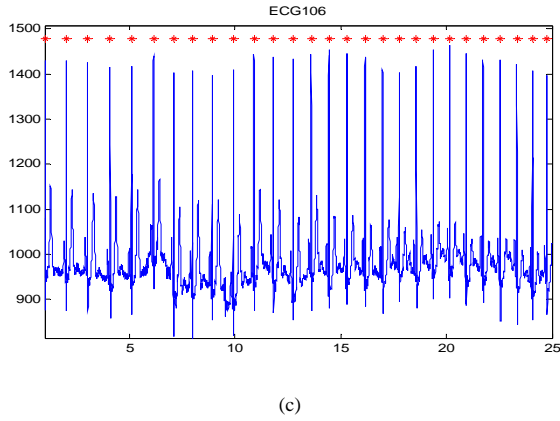


Fig. 5 Positioned R-peaks in ECG signal (a) ECG100, (b) ECG101, (c) ECG106.

Fig. 6 shows some examples of ECG signal (in blue color) and the Positioned P, Q, R, S, and T peaks (in red color).

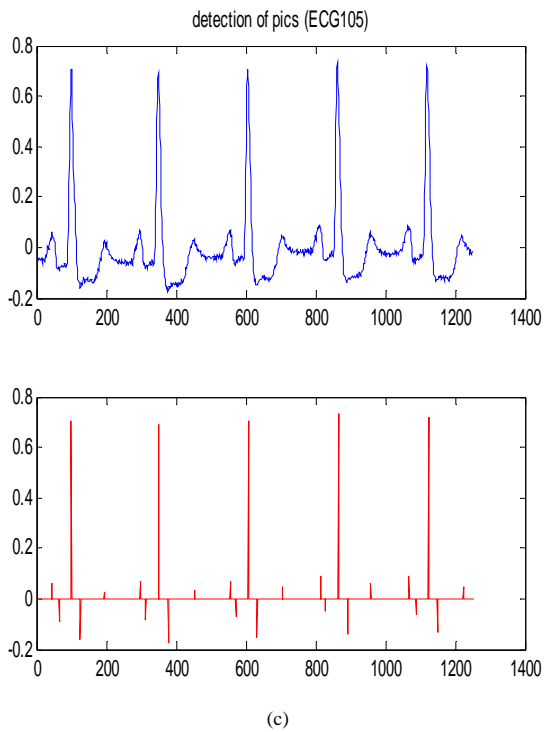
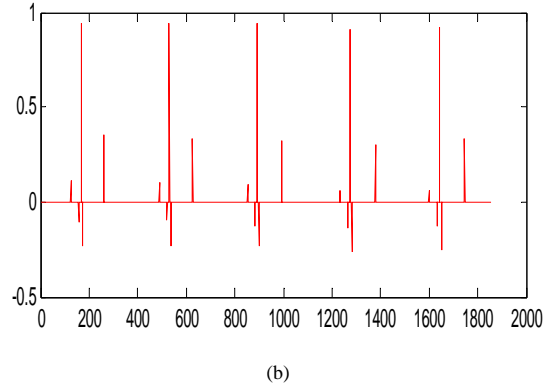
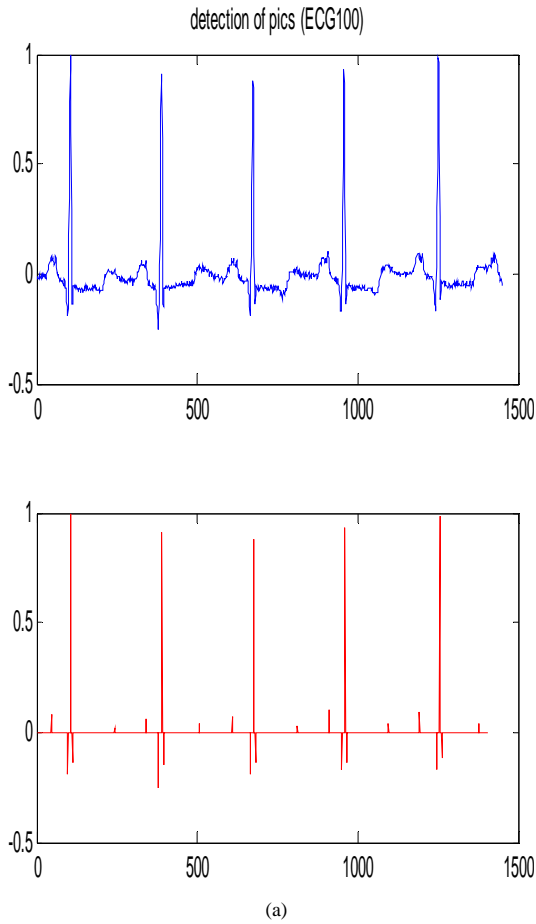


Fig. 6. Positioned P, Q, R, S, and T peaks in ECG signal (a)ECG100, (b)ECG106, (c)ECG105.



The proposed algorithm detection achieves very good detection performance. This algorithm attains sensitivity (Se) of 99.94 % and a positive predictivity (P+) of 99.94%. The sensitivity Se is defined as the probability that a sick patient to be detected:

$$S_e = \frac{TP}{TP+FN} \quad (7)$$

The positive predictivity is given by:

$$P^+ = \frac{TP}{TP+FP} \quad (8)$$

where TP is the number of beats correctly identified, FN the number of false detections, and FP the number of false positive misdetections.

Table 1 reports the obtained results from Sensitivity (Se) and positive predictivity (P+) computation by using the proposed detection technique.

Table 1: Sensitivity (Se) and Positive Predictivity (P+) Results obtained from the proposed detection technique application. Signal Records are chosen form MIT-BIH Database.

Tape (N°)	Total N° beats	FP beats	FN beats	P+ (%)	Se (%)
100	2273	0	0	100	100
101	1865	0	0	100	100
102	2187	0	0	100	100
103	2084	0	0	100	100
104	2230	1	0	99.95	100
105	2572	0	0	100	100
106	2027	0	0	100	100
107	2137	0	1	100	99.95
108	1763	13	21	99.26	98.82
109	2563	0	0	100	100
111	2124	0	0	100	100
112	2539	0	0	100	100
113	1795	0	0	100	100
114	1879	0	0	100	100

115	1953	0	0	100	100
116	2412	0	0	100	100
117	1535	0	0	100	100
118	2275	0	0	100	100
119	1987	0	0	100	100
121	1863	3	0	99.83	100
122	2476	0	0	100	100
123	1518	0	0	100	100
124	1619	0	0	100	100
200	2601	0	4	100	99.84
201	1963	0	0	100	100
202	2136	0	0	100	100
203	2982	3	0	99.89	100
205	2656	0	0	100	100
208	2956	2	0	99.93	100
209	3004	0	0	100	100
210	2647	3	0	99.88	100
212	2748	0	0	100	100
213	3251	0	0	100	100
214	2262	0	0	100	100
215	3363	0	2	100	99.94
217	2208	0	0	100	100
219	2154	0	0	100	100
220	2048	0	0	100	100
221	2427	0	0	100	100
222	2484	7	0	99.71	100
223	2605	0	0	100	100
228	2053	24	19	98.84	99.08
230	2256	0	0	100	100
231	1886	0	0	100	100

Table 2 reports the results obtained from Sensitivity (Se), positive, predictivity (P+) and %error computation by using the proposed detection technique and some others techniques used for evaluation. %error is expressed as follow:

$$\%error = \frac{FP+FN}{Total\ beats} \quad (9)$$

Table. 2: R wave's detection results on MIT-BIH database.

QRS detector	Se %	P+ %	% error
Arzeno et al.[16]	99.29	99.24	1.47
	99.57	99.59	0.84
	98.07	99.18	2.75
Huabin and Jiankiang [17]	99.68	99.59	0.73
Josko [18]	99.86	99.91	0.23
Mahmoodabadi et al.[1]	99.18	98	2.82
This work	99.94	99.94	0.12

Those results show clearly that the proposed method outperforms some conventional techniques used in our evaluation.

## 5. Conclusion

In this paper, we have developed a new method for R-peaks detection using the modified continuous wavelet transform (MMycwt) which is obtained from (Mycwt) used for the bionic wavelet transform (BWT). After detecting the R-peaks, Q and S are detected then the Modulus Maxima is applied to multi-scale product in order to detect P and T waves. This detection is performed frame by frame and each frame is localized between  $S_i$ -peak and  $Q_{i+1}$ -peak ( $S_i Q_{i+1}$ ). The multi-scale product is calculated from the product of undecimated wavelet coefficients of successive scales (scale1, scale2 and scale3). The undecimated wavelet coefficients are obtained from Undecimated Discret Wavelet Transform (UDWT) application to each modified frame. For evaluating the proposed method, we have used 46 half-hour recording for a total 23 hours of ECG data, extracted from MIT-BIH arrhythmia database. The advantages of this algorithm are; very fast to implement, easy to execute, and achieves very good detection performance. This algorithm attains  $Se=99.94\%$  and  $P+=99.94\%$ .

## References

[1] S.Z.Mahmoodabadi, A.Ahmadian, and M.D Abdolhasani, "ECG Feature Extraction Using Daubechies Wavelets", Proceedings of the Fifth IASTED International Conference VISUALIZATION, IMAGING, AND IMAGE PROCESSING (VIIP '05), 2005, pp. 343-348, Benidorm, Spain.

[2] Robert J. Huzar, "Basic Dysrhythmias Interpretation and Management", C.V. Mosby Co., 1988.

[3] Mneimneh MA, Corliss GF, Povinelli RJ, "A cardiac electro-physiological model based approach for filtering high frequency ECG noise", Computers in Cardiology, 34, 2007, pp. 109-112.

[4] LI Yan-jun, Yan Hong, WANG Zeng-li.. "A Comparative Study on Removal Methods of ECG Baseline Wandering", Space Medicine and Medical Engineering, 2009-05.

[5] I. Daubechies, "Ten lectures on Wavelets", Philadelphia, Society for industrial and applied Mathematics, 1992.

[6] B. Walczak, "Wavelets in Chemistry", The Netherlands, Elsevier Science, Data Handling in Science and Technology, Volume 22, 2000.

[7] Teolis, A., "Computational signal processing with wavelets", Springer, Birkhäuser Engineering, 1998.

[8] Samar Krimi, kais oui, and Nouredine Ellouze, "T-Wave Detection Based on an An Adjusted Wavelet Transform Modulus Maxima", International Journal of Biological and Life Sciences, v1, 2005, pp. 128-132.

[9] <http://www.physionet.org/physiobank/database/SVdb/MIT-BIH>.

[10] Felipe E.Olevera, Jr., and Student Member, IEEE, "Electrocardiogram Wave Feature Extraction Using the Matched Filter", ECE 510: STATISTICAL SIGNAL PROCESSING II, 2006, pp. 1-6.

[11] Wavelet denoising procedure in matlab, <http://www-ljk.imag.fr/SMS/software/GaussianWaveDen/>.

[12] O. Sayadi and M. B. Shamsollahi, Multiadaptive "Bionic Wavelet Transform: Application to ECG Denoising and Baseline Wandering Reduction", EURASIP Journal on Advances in Signal Processing, vol. 2007, pp. Article ID 41274, 2007, 11 pages..

[13] X. Yuan, "Auditory Model-based Bionic Wavelet Transform for Speech Enhancement", Ph.D. thesis, Lab Milwaukee, Marquette University, City, Wisconsin, May 2003.

[14] Johnson, M.T., Yuan, X. and Ren, Y., "Speech signal enhancement through adaptive wavelet thresholding", Science Direct, Speech Communication, Vol. 49, 2007, pp.123-133.

[15] R. Bessrou, Z. Lachiri, N. Ellouze, "UsingMultiscale Product for ECG Characterization", Hindawi Publishing Corporation Research Letters in Signal Processing, 2009, Volume 2009.

[16] N. M. Arzeno, Z.-D. Deng, and C.-S. Poon, "Analysis of first derivative based QRS detection algorithms," IEEE Transactions on Biomedical Engineering, vol. 55, no. 2, 2008, pp. 478-484.

[17] Z. Huabin and W. Jiankang, "Real-time QRS detection method," in Proceedings of the 10th International Conference on E-Health Networking, Applications and Services, July 2008, pp. 169-170..

[18] A. Josko, "Discrete wavelet transform in automatic ECG signal analysis," in IEEE Instrumentation and Measurement Technology Conference, 2007, Warsaw, Poland, 2007.

**First Author** Mourad Talbi received his Bachelor in Mathematics in 1997 from the Sciences Faculty of Tunis and he obtained his Master Degree at 2004 from the National School of Engineers of Tunis (ENIT), Tunisia, in Automatic and Signal Processing and he has obtained his

PhD Thesis at 2010 from Faculty of Sciences of Tunis, Tunisia, in Electronics. He is actually an assistant professor in high school of applied mathematics and physics of Kairouan and he is a Researcher Member of the Signal Processing Laboratory of Faculty of Sciences of Tunis, Tunisia.

**Second Author** Aouinet Akram received his master Diploma in electronics from Sciences Faculty of Tunis in 2010.

**Third Author** Riadh Baazaoui received his Bachelor in 2000 in mathematics from faculty of sciences of Tunis and his master Diploma in applied mathematics in 2003 from Sciences Faculty of Tunis. He is actually in PhD thesis of stochastic analysis particularly the applied mathematics of finance and physics.

**Fourth Author** Cherif Adnane obtained his Engineering Diploma in 1988 from the National School of Engineers of Tunis, and his PhD in Electrical Engineering and Electronics in 1997. Actually, he is a Professor at Science Faculty of Tunis, responsible for the signal processing laboratory. He participated in several research and cooperation projects, and is the author of many international communications and publications.

## **Presentation an approach for scientific workflow distribution on cloud computing data center's servers to optimization usage of computational resource**

Ahmad Faraahi<sup>1</sup>, Rohollah Esmaeli Manesh<sup>2</sup>, Ahmad Zareie<sup>3</sup>, Mir Mohsen Pedram<sup>4</sup>, Meysam Khosravi<sup>5</sup>

<sup>1,2</sup>Department of IT and Communication, Payam Noor University, Tehran, Iran.

<sup>3</sup>Young Researchers Club, songhor Branch, Islamic Azad University, songhor, kermanshah, Iran.

<sup>4</sup>Engineering Department, Faculty of Engineering, Tarbiat Moallem University, Karaj/Tehran, Iran

<sup>5</sup>Imam Reza Center Of Applied Science And Technology (IRCAST) Kermanshah, Iran.

### **Abstract**

*In a distributed system Timing and mapping the priority of tasks among processors is one of the issues attracted most of attention to itself. This issue consists of mapping a DAG with a set of tasks on a number of parallel processors and its purpose is allocating tasks to the available processors in order to satisfy the needs of priority and decency of tasks, and also to minimize the duration time of execution in total graph . Cloud computing system is one of the distributed systems which supplied a suitable condition to execute these kinds of applications, but according to the payment based on the rate of usage in this system, we should also consider reducing computation costs as the other purpose. In this article, we'll represent an approach to reduce the computation costs and to increase the profitability of computation power in Cloud computing system. To some possible extent, simulations show that the represented approach decreases the costs of using computation resources.*

### **Introduction**

*The timing issue of a graph, the functions of a parallel plan on the distributed computing system, is a Np-Complete issue and as one of the challenges in parallel computing problems has aroused a lot of attention [1]. Scheduling and timing of task's graph is an available process for mapping effective tasks on the number of processors that could be done considering purposes such as minimizing the application of execution time, increasing confidentiality, increasing resources profitability, decreasing data transmission costs and using saving resources and etc. [2]. Therefore, the efficiency of distributed system greatly*

*depends on the way of timing and task allocation method on them. In one hand, delays of communicative links among processors and their weak bandwidth among them may challenge task's graph application on parallel processors. One of the computing systems that are represented recently is Cloud computing system. The concept of this system has represented since 2007[7] and was successful in many field [3, 4, 5, 6]. The most important purpose of this system is decreasing economical costs and offering computing service such as water, electricity, gas and telephone. I.e. users use resources and services when they need according to the amount of their needs and also pay based on the rate of their use [8]. This system consists of many data centers that are distributed geographically all over the world and are accessible using internet. Each data center consists of many computing and saving servers and other resources. Servers in each center are attached with a high bandwidth in every spot and we can consider zero as their communicative delay among them.*

*According to the stated features cloud computing system is suitable condition for applying priority task's graph and while mapping graph on inter-center computing servers, it is possible to benefit a wide range of advantages such as, removing communicative delays using powerful saving computation servers, decreasing additional costs, profitability of high scaling cloud computing system and etc. But while mapping the parallel application on cloud computing system, some computing servers may remain unused effectively, and their profitability decreases to some extent. Executing task's graph on some processors always is not monotonous, since according to the parallel nature of tasks it might need four computing server in a moment and after completing tasks the number of required servers decrease to three server, these changes will be continued till the complementation of all graph. As in cloud computing system there is another simultaneous requests in which processors might serve them, therefore consecutive switches of computing servers are needed between simultaneous requests of other users and a user that represented the task graphs to the cloud computing system. In addition to decreasing profitability of computing resources of these switches, they may not be possible sometimes for severe oscillations,*

because based on the rate of overload that will have, it is better instead of switching to another use, let the computing server idle for a moment and restart executing other tasks of graph. To understand this issue clearly pays attention to the following example:

If task graph be like the fig.1

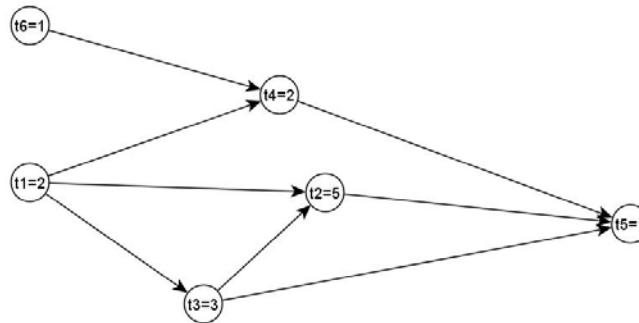


Fig 1- a sample of tasks graph

Two different mapping states are shown in figures 2 and 3 on two computing server for this graph:

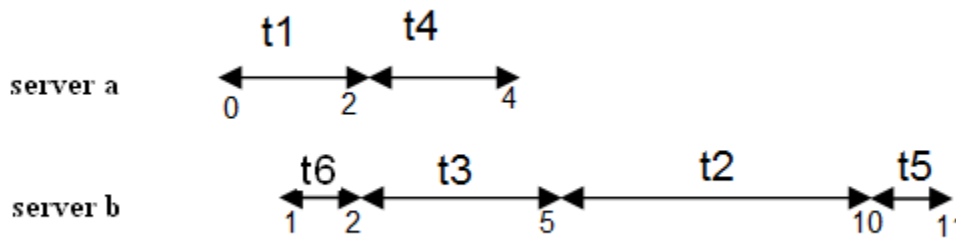


Fig 2- First mapping for graph in the figure 1

And second mapping for this graph:

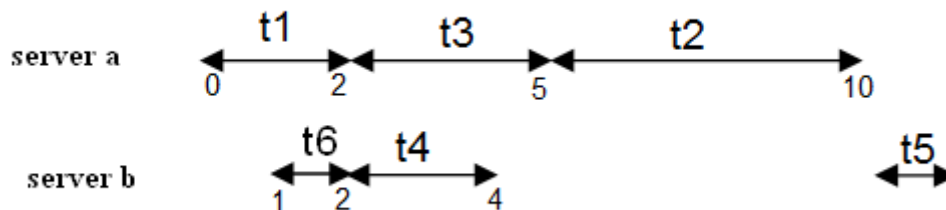


Fig 3- Second mapping for graph figure1

*As it is being observed in the first mapping server 'a' will be released in the moment of '4' that in the interval '0 to 4' completely uses computing server and also server 'b' that from moment 1 to 11 serve the graph is completely profitable, but in the second mapping server 'b' from moment 1 to 11 serve the graph and from the moment 4 to 10 it is useless and idle, so in some extent its profitability decreases because if in this interval the processor execute the other users 'simultaneous requests and again return to the application based on overload it is inefficient and it is of high costs. In this article we want to represent an approach for mapping task's graph on computing servers in a data center of cloud computing system in order to make profit from available servers and to decrease its idle and useless time.*

### **Related works**

Timing and mapping are one of the main challenges in executing the parallel applications on distributed systems and aroused a lot of attention. To map and schedule task graph various approaches are presented that in each one some limitations are considered for system and timing has been done schedule has been done for different purposes. This approach can be divided into the following classes [28, 34]:

- Task Duplication Based (TDB) [9][10][11]
- Bound Number of Processors (BNP) Scheduling [12][13][14][15]
- 
- Unbounded Number of Clusters (UNC) Scheduling [15][16]
- Arbitrary Processor Network (APN) scheduling [17][18][19][20]
- List scheduling algorithms [35][36][37][38][39]

*In the Task Duplication Based (TDB) scheduling with the purpose of decreasing Schedule length some tasks duplicated and run on several servers, but using this technique in cloud computing system is inefficient and leads to increase in computation costs. Some algorithms suppose that there exist an unlimited number of processors to apply tasks. But strategies of second class schedule consider a high limit for processors that are available. According to the limitation*

*of computing servers in the centers of cloud computing systems a high limitation was considered for processors. Schedule algorithms with unlimited number of clusters, in the beginning of schedule process each node is considered as a cluster and in the next step when two clusters combination decrease the completion time, two cluster combine and this combination continue recessively. The reason for this kind of algorithms is using plenty of processors to decrease the application time. Therefore, this approach needs another stage for clusters mapping on processors, because available processors might be less than clusters.*

*Another important feature in schedule is considering the type of server attachment and network topology. As servers in a cloud computing data center are attached they completely connected with a high bandwidth, in our work, topology is considered completely attached. Of the other represented schedule we can refer to following cases.*

*Kumar represented an approach for scheduling independent works in a distributed system that the only relationship between tasks is the need of reaching shared files and no task is needed for the results of the other application task [21]. In [22] a schedule on grid system that is represented data replacement activities are closely similar to computing tasks and they can be queuing, scheduling and managing. In this article [23] also a data placement approach represented in the cloud environment and the only task placement criteria for place of processing tasks has set a center with tasks data and the short length of schedule is not considered.*

*In distributed system and in the field of task designation the other kinds of timing with different purposes are represented. In [25], genetic algorithm is used for mapping a task graph in the distributed systems to decrease the length of schedule and load balancing on the servers. In [9], an approach to scheduling tasks has been stated for the purpose of fulfillment of task completion deadlines and optimization of execution costs. in approach [21] schedule tasks in order to decrease schedule length and increasing reliability of executing application in distributed environment heterogeneous . This Article [28] represented a method for scheduling the*



*workflow on processor with various abilities. Of the other approaches that are represented for allocation of independent tasks in homogenous distributed environment to increase reliabilities we can refer to [29, 30]. In the previous work [31] we represented an approach for scheduling a work flow graph. In this approach we tried to decrease storage costs, as possible as, during the executing the graph. In this work we want to represent an approach in order to increase profitability and usage of computing servers.*

### **Stating the main problem**

In this article we map the workflow of graph on some computing servers in a data center, which are servers completely connected. The number of servers are shown using  $M$  such as  $m_1, m_2, \dots, m_M$ .

The workflow graph of DAG consists of  $N$  nodes that we have  $n_1, n_2, \dots, n_N$  that each node shows just one task of work graph. Each task consists of a set of instructions that should be executed on a computing server without retaking and respectively. Each node has one or more parent task and when it starts running that running of all parent tasks being ended. In this article we call a node without parent 'fn' and the node without child 'ln'. Also graph has the  $E$  directed edge where shows the priority and the consequence of running tasks. It means  $n_i \rightarrow n_j$  edge show that  $n_j$  is a child for  $n_i$  and not until the end of  $n_i$ ,  $n_j$  can not start. In some researches usually there is a weight for edges that show communicative cost or duration till output data of  $n_i$  are given to server  $n_j$  (if  $n_i$  and  $n_j$  are mapped in different centers). But as we map graph on computing servers in a center and communicative links between servers in a data center have a high bandwidth. Communicative costs from the time giving  $n_i$  data to  $n_j$  is zero, therefore we consider no weight for edges. If mapping  $n_i$  on server  $m_j$ , the beginning time will be shown by  $ST(n_i, m_j)$ , and its completion time is shown by  $ET(n_i, m_j)$ . In which according to being homogeneous computing servers:

$$ET(n_i, m_j) = ST(n_i, m_j) + w(n_i)$$

That  $w(n_i)$  shows the task of executing time. Our purpose in this work is increasing profitability and decreasing the idle time of computing servers, it

means that when two tasks of  $n_i$ ,  $n_t$  are mapped as two consecutive tasks on server  $m_j$  the interval between completion  $n_i$  and beginning  $n_t$  means that:

$$ST(n_i, m_j) - ET(n_t, m_j)$$

It should be somehow that be able to use server  $m_j$  to process simultaneous requests of other users, if this amount is small enough for processing other requests to use, server should be idle in this time interval and this lead to decline in server profitability.

So if

$$0 < ST(n_i, m_j) - ET(n_t, m_j) < \lambda \quad (2)$$

Computing server could not be used in the time interval between these two tasks, therefore it will be useless for sometimes. The value of  $\lambda$  should be considered according to the application time for works as well, that in recent  $c$  request are sent to consider data center. So, if in equation 2 is not appointed we can retake computing server from the workflow graph and use in the time interval between beginning and completion of tasks and set them for applying other requests, but in spite of this according to the overload that lead to computing server switch, it is better that intended server stay for work graph user and in this time stay useless and free.

Therefore when work graph usage has the computing server will be calculated by equation 3.

$$\begin{aligned} \text{Allocate } (m_j) &= w(n_{j1}) + \sum_{k=2}^l w(n_{jk}) + \alpha & (3) \\ \text{IF } ST(n_{jk+1}, m_j) - ET(n_{jk}, m_j) < \lambda & \text{ then } \alpha = ST(n_{jk+1}, m_j) - \\ ET(n_{jk}, m_j) & \text{ else } \alpha = 0 \end{aligned}$$

That  $w(n_{jk})$  intention the time of  $k$ th task that mapped on  $m_j$  computing server and  $l$  is the number of these tasks.

And the time that  $m_j$  computing server is to serve workflow graph usage will be calculated using the relation 4.

$$\text{Process } (m_j) = \sum_{k=1}^l w(n_{jk}) \quad (4)$$

Therefore profitability rate of computing server  $m_j$  will be obtained from relation 5.

$$\text{useful}(m_j) = \frac{\text{Process}(m_j)}{\text{Allocate}(m_j)} \quad (5)$$

And the rate of profitability from all of the calculating servers that serve workflow graph usage application will be obtained from relation 6.

$$\text{Useful\_Allservers} = \sum_{k=1}^M \text{Useful}(m_k) \quad (6)$$

Therefore, the purpose of represented approach is defined as follow:

Max(Useful All servers)

In other hand, the application of whole graph should be ended in an acceptable time; it means that, if possible, schedule length or  $n_N$  completion time might be minimized.

### Presented approach

Before stating represented approach it is necessary to define two terms. As for each task till not completion of all parents the possibility of starting application is not supplied, therefore the earliest starting time for task  $t_i$  is shown by  $es(t_i)$  and is defined as follow :

$$\begin{cases} \max_{i \in \text{pred}(t_i)} (et(t_i)) & \text{else} \\ 0 & \text{if } i = 1 \end{cases}$$

Then we express a suggested approach.

For each pair of work  $(t_i, t_j)$  we calculate consistency rate based on the following relation.

$$\text{consistency}(t_i, t_j) = \begin{cases} 1 - \frac{|St(t_i) - Et(t_i)|}{St(t_N) - Et(t_1)} & \text{else} \\ 1 & St(t_j) - Et(t_i) \geq \lambda \end{cases}$$

Then we create an  $n \times n$  matrix named AM that we set the element  $AM_{i,j}$  is  $\text{consistency}(t_i, t_j)$  This matrix will be a polar matrix.

Then using BEA ( Bond Energy Algorithm)we make some changes on matrix. In 1972 this algorithm [32] in the system of distributed information bank is used for vertical sectioning the great tables [33].

This algorithm is a transformational algorithm that with transformations on columns and rows, categorize elements that are more consistent.

In this work we give AM matrix to BEA algorithm as entrance. After doing xxxfollowing relation will be selected as breaking point and then we converse the matrix to two sub-matrixes. By this break, most consistent elements will be categorized in one group.

$$\text{break point} = \max \left[ \sum_{i=1}^p \sum_{j=1}^n AM_{ij} \times \sum_{i=p+1}^n \sum_{j=p+1}^n AM_{ij} - \left( \sum_{i=1}^p \sum_{j=p+1}^n AM_{ij} \right)^2 \right]$$

That in the above relation for  $p=1,2,\dots,n-1$  find the maximum amount and we divide the matrix into two groups. Now for each matrix we calculate break point recessively and the sub-matrix that has the highest break point value has more inconsistent elements and is selected to break it into two sub-matrixes and will be sent to BEA algorithm as entrance data. The stated process is done recessively till the number of sub-matrixes (groups) reach to the number of servers (M) or break point for all sub-matrixes equals to zero (it means that all of the elements in a matrix from all sub-matrixes are completely consistent) then works of each group are arranged based on the most early possible time to start (ES) in an ascending form and each of these groups could be selected for executing on one of the computing servers.

### Experiments and Assessment of the represented approach

To evaluate the efficiency of represented approach we manipulate their codes in C# environment and apply them on different graphs with different parameters and in each state the completion workflow time and profitability rate from available servers are calculated and then we show the results using Matlab software in the form of graphs. For analyzing the amount of success in the represented approach we compare the result with schedule approaches based

on scheduling list. In this represented approach the main idea is allocating the priority to each task and creating an ascending list of tasks based on priorities, then not until losing all elements in the list in the beginning of the list a deletion will be mapped on a free server [28].to contrast the completion time of workflow we used a normal schedule length that is defined as following relation [30].

$$NSL = \frac{eT_{\Sigma}}{\sum_{T_i \in CP} W(T_i)}$$

And also to contrast profitability rate of computing servers we used the following relation:

$$Useful\_Rate = Useful\_AllServer * 100$$

We have applied both approaches on graphs with the same parameters and also in each experiment we evaluated the effects of changing one parameter from workflow graph on approaches. All of the data in graphs is the average of the results getting from the applied approaches on 50 different entrance graphs with the same parameters.

### Experiment 1

In this experiment we analyze changing the effect of graph tasks on approaches. The results of this experiment are shown in figure 4.

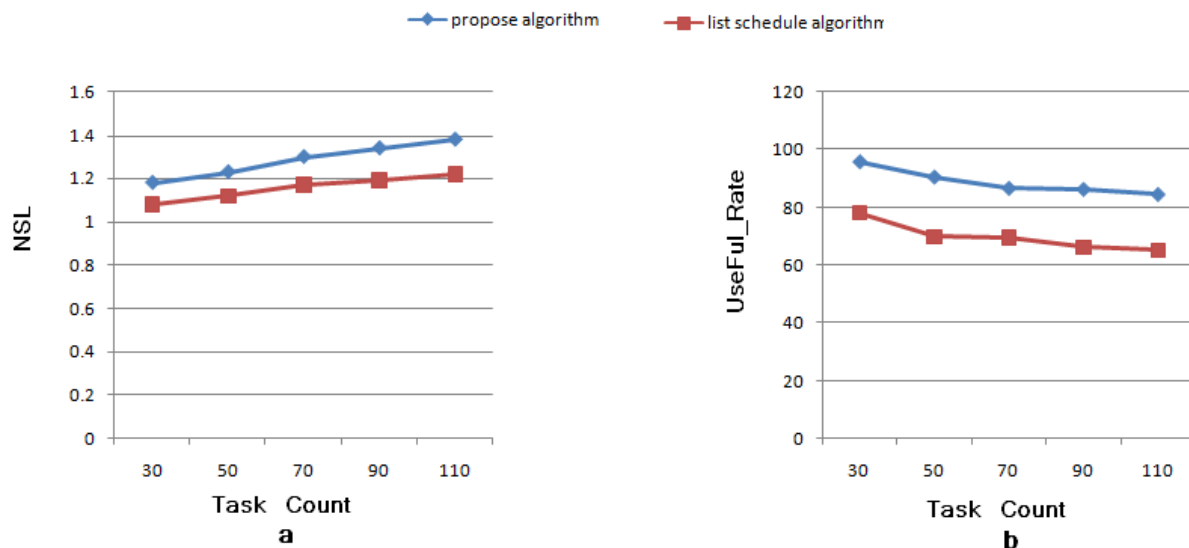


Fig 4- comparing two approaches based on changing many tasks of workflow

according to the results getting from this experiment in this represented approach the profitability of servers are more than the schedule list and by increasing the number of tasks the represented approach may have a better result, because this approach try to group the most consistent tasks, so by increasing the number of tasks it is possible to find the most consistent works in the graph and to map them on the server. This issue increases profitability of server. But according to the figure 4-a length of schedule will be increased in this approach.

### Experiment2

In the second experiment we have applied algorithms on graphs with the same parameters and also we have shown the results of two algorithms affected by changing communicative coefficient among tasks. Figure 5 shows the useful rate and NSL for two algorithms.

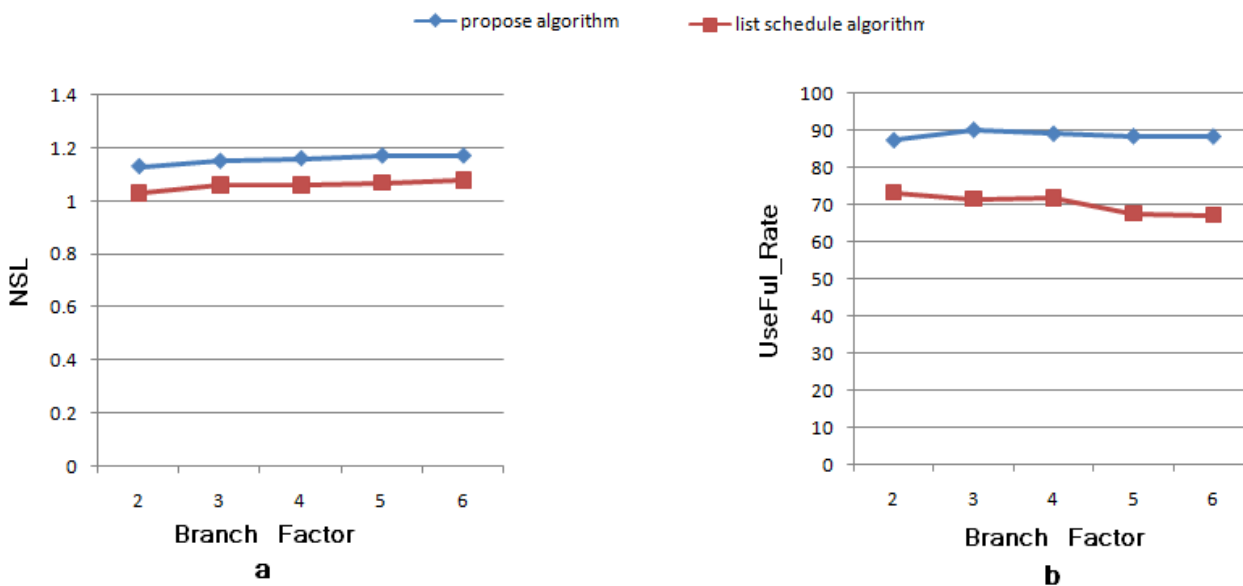


Fig 5- Comparing two approaches effected by changing coefficient of communication among workflow graph tasks

According to the results getting from the second experiment, the represented approach shows stable results by increasing communicative coefficient among tasks, but in the schedule list approach the profitability of servers will be decreased by increasing communicative coefficient and also because of the interval for applying tasks on each server. But according to this experiment the schedule duration is greater in the represented approach.

### Experiment 3

In this experiment we had applied algorithms on graphs with the same parameters on different number of servers. We have shown the useful rate and NSL resulted from two algorithms.

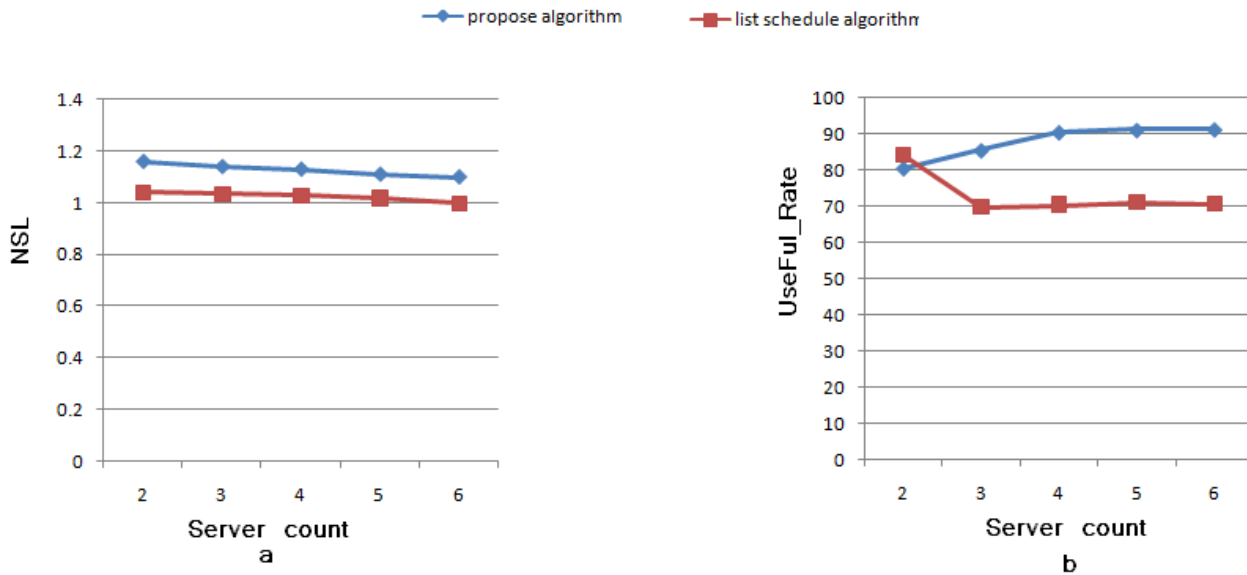


Fig 6- Analyzing efficiency of two approaches based on changing the number of servers.

According to the results of third experiment by increasing the number of servers, the represented approach looking for better and more effective use of them because by increasing the number of servers sub-matrixes are increasing, so the tasks in each matrix have higher consistency and this issue result in a higher profitability of servers. Also based on figure 6-a by increasing the number of servers, schedule length will decrease in both algorithms, but we can see a higher schedule length in represented approach again.

### Results and future works:

Mapping and distributing trains of a parallel application tasks have been expressed as one of the challenges of distributed systems. According to the paying computing costs, saving, network and ...based on the rate of usage, geographical domain and cloud computing system- based on internet and also simultaneous requests from different users, algorithms that are used to map applications in this system should prevent from being uselessness. the

advantages of this issue are increasing the usage efficiency of computing servers, decreasing payment costs to the clients of the cloud computing system services, increasing the ability of cloud computing service suppliers in response to simultaneous requests from the other clients, decreasing workflow completion time that lead to reducing clients patience.

In addition, in scheduling application approach we should schedule them in order to minimize the whole completion time of application.

Also, this issue leads to decrease in using computing services and save them so that finally lead to consumers and suppliers satisfaction of the cloud computing services. Simultaneous realizing these two purposes might be opposite optimally. So in this research we represented an approach that ensures getting to these two purposes. The results of experiments show that the represented approach is successful in this way. In this work, we did not use work and data propagation techniques, because according to the payment and based on the rate of using in cloud system, applying this technique lead to the increasing in costs.

Other main challenges in cloud computing system are reliability of resources during the application of parallel uses. In next works we try to represent an approach that produces further data during workflow of application. Since another tasks are mapped on servers with higher reliability. When the computing server damages, this issue causes suspending fewer tasks than workflow application and application will be completed with less delays.

## Resource

- [1] H. El-Rewini, T.G. Lewis and H.H. Ali, Task Scheduling in Parallel and Distributed Systems, Prentice-Hall International Editions (1994).
- [2] M. Rahman, R. Ranjan and R. Buyya, Cooperative and decentralized workflow scheduling in global grids, *Future Generation Comp. Syst.* 26(5) (2010), PP 753-768.
- [3] M. Brantner, D. Florescu, D. Graf, D. Kossmann, T. Kraska, Building a Database on S3, in: SIGMOD, Vancouver, BC, Canada, 2008, pp. 251–263.
- [4] R. Grossman, Y. Gu, Data Mining Using High Performance Data Clouds: Experimental Studies Using Sector and Sphere, in: SIGKDD, 2008, pp. 920–927.
- [5] R. Buyya, C.S. Yeo, S. Venugopal, Market-oriented cloud computing: Vision, hype, and reality for delivering IT services as computing utilities, in: 10th IEEE International



Conference on High Performance Computing and Communications, HPCC-08, Los Alamitos, CA, USA, 2008.

[6] C. Moretti, J. Bulosan, D. Thain, P.J. Flynn, All-Pairs: An abstraction for data-intensive cloud computing, in: *IEEE International Parallel & Distributed Processing Symposium, IPDPS'08*, 2008, pp. 1–11.

[7] A. Weiss, Computing in the Cloud, vol. 11, *ACM Networker* (2007) 18–25.

[8] R. Buyya, C. Yeo, S. Venugopal, J. Broberg and I. Brandic, Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility, *Future Generation Computer Systems*, 25(6) (2009), pp. 599–616.

[9] I. Ahmad and Y.-K. Kwok, On exploiting task duplication in parallel program scheduling, *IEEE Trans. Parallel Distrib. Syst.* 9 (9) (1998), pp. 872–892.

[10] S. Darbha and D.P. Agrawal, Optimal scheduling algorithm for distributed - memory machines, *IEEE Trans. Parallel Distrib. Syst.* 9 (1) (1998), pp. 87–95.

[11] C.H. Papadimitriou and M. Yannakakis, Towards an architecture-independent analysis of parallel algorithms, *SIAM J. Comput.* 19 (2) (1990), pp. 322–328.

[12] M.K. Dhodhi, I. Ahmad and A. Yatama *et al.*, An integrated technique for task matching and scheduling onto distributed heterogeneous computing system, *J. Parallel Distrib. Comput.* 62 (9) (2002), pp. 1338–1361.

[13] H.J. Park and B.K. Kim, An optimal scheduling algorithm for minimizing the computing period of cyclic synchronous tasks on multiprocessors, *J. Syst. Softw.* 56 (3) (2001), pp. 213–229.

[14] Xiaoyong Tang, Kenli Li, Guiping Liao and Renfa Li, List scheduling with duplication for heterogeneous computing systems, *J. Parallel Distrib. Comput.* 70 (4) (2010), pp. 323–329.

[15] D. Kim and B.G. Yi, A two-pass scheduling algorithm for parallel programs, *Parallel Comput.* 20 (6) (1994), pp. 869–885.

[16] Y.-K. Kwok and I. Ahmad, Dynamic critical-path scheduling: an effective technique for allocating task graphs onto multiprocessors, *IEEE Trans. Parallel Distrib. Syst.* 7 (5) (1996), pp. 506–521.

[17] S. Bansal, P. Kumar and K. Singh, An improved duplication strategy for scheduling precedence constrained graphs in multiprocessor systems, *IEEE Trans. Parallel Distrib. Syst.* 14 (6) (2003), pp. 533–544.

[18] G.C. Sih and E.A. Lee, A compile-time scheduling heuristic for interconnection-constrained heterogeneous machine architectures, *IEEE Trans. Parallel Distrib. Syst.* 4 (2) (1993), pp. 175–187.

[19] Xiaoyong Tang, Li Kenli and D. Padua, Communication contention in APN list scheduling algorithm, *Sci. China Ser. F* 52 (1) (2009), pp. 59–69.

[20] H. El-Rewini and T.G. Lewis, Scheduling parallel program tasks onto arbitrary target machines, *J. Parallel Distrib. Comput.* 9 (2) (1990), pp. 138–153.

- [21] K. Kaya, B. Uçar and C. Aykanat, Heuristics for scheduling file-sharing tasks on heterogeneous systems with distributed repositories, *J. Parallel Distrib. Comput.* 67 (2007), pp. 271 – 285.
- [22] T. Kosar, M. Livny, Stork: Making data placement a first class citizen in the grid, in: Proceedings of 24th International Conference on Distributed Computing Systems, ICDCS (2004), pp. 342–349.
- [23] D. Yuan, Y. Yang, X. Liu, J. Chen, A Data Placement Strategy in Cloud Scientific Workflows , *Future Generation Computer Systems (FGCS)*, 26(8)(2010), pp. 1200-1214.
- [24] T. Kosar, S. Son, G. Kola, M. Livny, Data placement in widely distributed environments , *Advances in Parallel Computing* 14(2005), pp. 105-128.
- [25] Fatma A. Omara, Mona M. Arafa, Genetic algorithms for task scheduling problem , *Journal of Parallel and Distributed Computing* 70(1)(2010), pp. 13-22.
- [26] Y. Yuan, X. Li, Q. Wang, X. Zhu, Deadline division-based heuristic for cost optimization in workflow scheduling , *Information Sciences* 179(15) (2009), pp. 2562-2575 .
- [27] X. Tang, K. Li, R. Li, B. Veeravalli, Reliability-aware scheduling strategy for heterogeneous distributed computing systems , *Journal of Parallel and Distributed Computing* 70(9)(2010), pp. 941-952 .
- [28] Z. Shi, Jack J. Dongarra, Scheduling workflow applications on processors with different capabilities , *Future Generation Computer Systems* 22(6)(2006), pp. 665-675.
- [29] Q. Kang, H. He, H. Song and R. Deng, Task allocation for maximizing reliability of distributed computing systems using honeybee mating optimization, [Journal of Systems and Software](#) 83(11) (2010), pp. 2165-2174.
- [30] C. Chiu, Y. Yeh, J. Chou, A fast algorithm for reliability-oriented task assignment in a distributed system, *Computer Communications* 25(17)(2002), pp. 1622-1630 .
- [31] A. zareie, M. M. pedram, M. kelarestaghi, A. kosari, *International Journal of Computational Intelligence and Information Security* 2(7)(2011), pp . 13-20.
- [32] W.T. McCormick, P.J. Schweitzer and T.W. White, Problem decomposition and data reorganization by a clustering technique, *Operations Research* 20 (1972), pp. 993–1009.
- [33] M.T. Ozsu and P. Valduriez, Principles of Distributed Database Systems, Prentice-Hall, Inc., Upper Saddle River, NJ, USA (1991).
- [34] X. Tang, K. Li, R. Li and B. Veeravalli , Reliability-aware scheduling strategy for heterogeneous distributed computing systems, [Journal of Parallel and Distributed Computing](#) 70(9)(2010), pp. 941-952.
- [35] M.Y. Wu and D.D. Gajski, Hypertool: A programming aid for message-passing systems, *IEEE Trans. Parallel Distrib. Syst.* 1 (1990) (3), pp. 330–343.
- [36] G.C. Sih and E.A. Lee, A compile-time scheduling heuristic for interconnection-constrained heterogeneous processor architectures, *IEEE Trans. Parallel Distrib. Syst.* 4 (1993) (2), pp. 175–187.
- [37] H. El-Rewini and T.G. Lewis, Scheduling parallel program tasks onto arbitrary target machines, *J. Parallel Distrib. Comput.* 9 (1990) (2), pp. 138–153.
- [38] G. Liu, K. Poh and M. Xie, Iterative list scheduling for heterogeneous computing, *J. Parallel Distrib. Comput.* 65 (2005) (5), pp. 654–665.

[39] R. Sakellariou, H. Zhao, A hybrid heuristic for dag scheduling on heterogeneous systems, in: Proceedings of 13th Heterogeneous Computing Workshop, HCW2004, Santa Fe, NM, 2004.

# Comparison of Routing Protocols for Locating moving object in Large Scale Cellular Wireless Sensor Network

Ola A. Al-Sonosy<sup>1</sup>, Mohammed A. Hashem<sup>2</sup> and Nagwa Badr<sup>3</sup>

<sup>1</sup>Information System Department, Faculty of Computer and Information Science, Ain Shams University, Cairo, Egypt

<sup>2</sup>Information System Department, Faculty of Computer and Information Science, Ain Shams University, Cairo, Egypt

<sup>3</sup>Information System Department, Faculty of Computer and Information Science, Ain Shams University, Cairo, Egypt

## Abstract

A Wireless Sensor Network (WSN) is ad hoc dense wireless network of small, low-cost sensor nodes (SN), which collect and disseminate sensed data with limited power sources. WSNs can be used in many applications, such as environmental monitoring and object tracking. Organizing network topology with energy efficient traffic can maximize network lifetime.

A cellular architecture based on organizing SNs according to their geographical position had been introduced. It uses inheritance features to organize and reduce traffic. It works with ALL-Active (AA) SNs. Moreover, two Energy aware models are suggested to prolong the WSN lifetime. The first uses Cellular Automata (CA) rules and the second uses Radio Triggered (RT) SNs.

Multi-hop, Static Tree, Aggregated Tree, Static cluster, and dynamic cluster based routing protocols has been designed and compared for AA, CA, and RT architectures resulting in longest network lifetime with no overhead communication for RT network with least energy consumption applying dynamic cluster routing.

**Keywords-** *Wireless Sensor Network, Cellular network, Ad-Hoc.*

## 1. Introduction

The development of large-scale sensor networks has drawn a lot of attention. Therefore, the large size of wireless sensor networks inevitably introduces significant scalability concerns. One of the main challenges is then to set up new architectures and mechanisms that can efficiently scale up with the growing number of nodes [1][2][3].

Large-scale WSNs consist of hundreds and thousands of nodes, and each node is able to sense the environment, perform simple computations and communicate with its other sensors or to the Based station, which is a central unit. The sensor node has qualities of low-cost, small size and power-saving. However, these qualities put bonds to the energy, storage capacity and computing ability. The most influential factor is the restriction to energy [4]. Achieving maximum lifetime in WSNs by optimally using the energy within sensor nodes has been the subject of significant researches in the last recent years. In this field, radio transmission and reception operations are being identified as one of the most energy consuming features.

Recently WSNs are often used to replace mankind with sensor nodes to obtain the information which is unable for human beings. Hence its applications is widespread, such fields in environmental, health, military. The military community is interested in deploying WSN to provide battlefield surveillance, reconnaissance, and battle damage assessment [5]. In these applications, WSN are deployed on-demand to monitor large terrains and obtain timely information about the enemy activities. Therefore, there is a need for a specially designed large scale WSN for object tracking.

The research here introduced large scale WSN specially designed for object tracking using acoustic sensors. Two energy aware architectures are also designed to prolong network lifetime. A design for Multi-hop, Static Tree, Aggregated Tree, Static cluster, and dynamic cluster based routing protocols is introduced and compared for the three network models.

The rest of the paper is organized as follows: section 2 is the previous work, section 3 contains the proposed cellular network architecture, section 4 includes two energy aware architectures, section 5 apply location of moving object to the suggested architectures with designs for the previously mentioned routing protocols, section 6 contains the simulator and its results and a comparison between the routing protocols through the networks, and section 7 concludes the paper.

## 2. Previous work

Moving object tracking using WSN has received considerable attention in recent years and intended solutions can be mainly classified into four schemes, which are: tree - based tracking, cluster-based tracking; prediction-based tracking; multicast message-based tracking methods[7].

In tree-based target tracking, nodes in a network may be organized in a hierarchical tree or represented as a graph in which vertices represent SNs and edges are links between nodes that can directly communicate with each other.

In the cluster architectures, clusters are formed statistically at the time of network deployment and the

properties of each cluster are fixed such as number of members, area covered. Static clustering is simple but has scalability problems. In contrast, dynamic clustering offers several advantages where clusters are formed dynamically depending on occurrence of certain events, thus redundancy data and interference is reduced [8].

Prediction-based tracking processes historical data depending on the object moving pattern to deduce subsequent movement of a mobile object [9].

Mobicast message-based tracking depends on prediction. It is a spatiotemporal multicast method in which message is delivered to a group of nodes that change with time according to estimated velocity of moving entity [10]. Some protocols employ clustering, prediction and scheduling mechanism thus giving better performance in terms of energy consumption as compared to other approaches but further metrics still need to be investigated [11].

### 3. Proposed Cellular Architecture:

**Topology Control Problem:** Topology Control Problem in WSN can be divided into two categories: Sensor Coverage Topology and Sensor Connectivity Topology. The coverage topology describes the topology of sensor coverage and is concerned about how to maximize a reliable sensing area while consuming less power. The connectivity topology on the other hand is more concern about network connectivity and emphasizes the message retrieve and delivery.

**Sensor Energy Consumption:** The energy consumed by the transmitting circuit is related with the distance transmitted the message [18]. The energy consumed by sending  $m$  bits message is calculated as follow:

$$E_{trans}(m,d) = \begin{cases} E_{elec} * m + E_{fs} * m * d^2 & \text{if } d < d_0 \\ E_{elec} * m + E_{mp} * m * d^4 & \text{if } d \geq d_0 \end{cases} \quad (1)$$

and to receive this message, the radio expends:

$$E_{rec}(m) = E_{elec} * m \quad (2)$$

$E_{trans}(m, d)$  is the energy consumed by sending  $m$  bites of messages.  $E_{rec}(m)$  is the energy consumed by receiving  $m$  bites of messages.  $d$  is the transmission distance.  $E_{elec}$  is the energy consumed by receiving or sending one bites of message.  $E_{fs}$  is the transmission constant of freedom space.  $E_{mp}$  is the comedown transmission constant of multi-path.  $E_{fs}$  and  $E_{mp}$  is related with the model of transmission channel used in the wireless sensor network.  $d_0$  is the maximum of transmission distance.

$$d_0 = \sqrt{\frac{E_{fs}}{E_{mp}}}$$

- **Network initialization and localization:**

Deterministic deployment of large WSNs is impractical due density required to provide appropriate network coverage. Furthermore, several applications of WSN are expected to operate in hostile environments [12]. Consequently, stochastic deployments become more feasible [13]. Stochastic deployment makes the network topology random. Since there is no a priori communication protocol, the network is ad hoc. According to the hostile environment in which the WSN can be deployed, increasing the number of sensors, to correlate the detection signals and so to maximize the probability of an accurate detection can be a solution. For these circumstances precise knowledge of node location in ad hoc sensor networks is an essential step in wireless networking. Unfortunately, for a large number of SNs, straightforward solution of adding GPS to all nodes in the network is not feasible due to the high cost of GPS.

Consider the case when we have deployed a sensor network consist of  $n$  sensors, with some SNs equipped with GPS are installed in the field known as *anchors* or *beacons*. Initially, anchors are aware of their own positions. Then all the other nodes localize themselves with the help of location references received from the anchors using the hyperbolic trilateration method introduced in [14]. This process ends with the result that all the SNs in the network knows its own position  $N(x,y)$ .

**Cellular layer deployment:** Communication cost in WSNs is at least two orders of magnitude higher than computation costs in terms of consumed power [15]. This communication-computation trade-off is the core idea behind low energy sensor networks. The suggested architecture applies a virtual cellular layer deployment over the unstructured network area as shown in Fig. 1 to organize and reduce communication through the network.

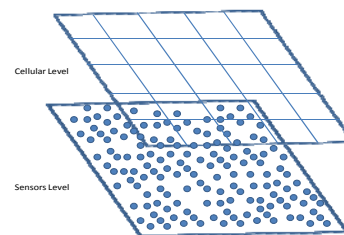


Fig. 1 Cellular-layer deployment

The network suggested architecture is as follows:

- **Wireless sensor network environment:**

- Assuming an area that we will deploy sensor nodes in as a coordinate known area and we will denote it as network area.

- The system has a single base station (BS) located at the center of that network area in order to be closer to all sensors in the network.
- Sensors deployed in the network area are assumed to be uniformly and densely distributed over the network area and initially have equal battery powers.
- As shown previously in the equation above, when transmitting a packet distance less than do, the sending energy consumed proportional to square the transmitted distance. Also the equation shows that, when transmitting a packet distance exceeds do, the sending energy consumed proportional to quadratic the transmitted distance. Therefore, in order not to increase the energy consumption of sending packets, we say that the **communication range** is the circle surrounding the sensor node that its center is the sensor node and its radius is  $R_{com}$  equals to do.
- The sensor node's **sensing range** is assumed to be as circle its center is the sensor node and its radius is  $R_{sense}$
- The sensing range of a sensor node is assumed to be half its communication range. Therefore  $R_{com} = 2 R_{sense}$ .
- **Cells Definition:** The network area is divided into equal sized non-overlapping areas called cells. Each cell has a known coordinates, and its size is assumed to be less than the sensing range of the deployed sensors as shown in figure 3. All sensor nodes belonging to one cell are assumed to have a common sensing and communication ranges. The communication range of a cell can be defined as the intersection of the communication ranges of the four sensor nodes residing at the corners of the cell. Also the sensing range can be defined as the intersection of the sensing ranges of the four sensor nodes residing at the corners of the cell. Both ranges are shown in figure 3. Applying the localization algorithm, described in the previous section, each node will identify its coordinates, consequently identifying to which cell it belongs.

**Cell Identification:** Each cell is identified by a pair of numbers (x,y), as shown in figure 2 starting from the origin, cell(0,0) in which the based station is located, and outwards to include all the network area.

**The cell size:** The maximum cell diagonal can be identified as  $d_0/4$ , as shown in figure 4. Therefore each cell can be projected to network area as a  $S_{cell} * S_{cell}$  square, where

$$\text{The cell Diagonal is } d_0/4 = \sqrt{2}(S_{cell})^2$$

Thus the cell side can be calculated as

$$S_{cell} = \frac{d_0}{4\sqrt{2}}$$

$$S_{cell} = \frac{\sqrt{E_{fs}}}{4\sqrt{2 * E_{mp}}}$$

(-5,5)	(-4,5)	(-3,5)	(-2,5)	(-1,5)	(0,5)	(1,5)	(2,5)	(3,5)	(4,5)	(5,5)
(-5,4)	(-4,4)	(-3,4)	(-2,4)	(-1,4)	(0,4)	(1,4)	(2,4)	(3,4)	(4,4)	(5,4)
(-5,3)	(-4,3)	(-3,3)	(-2,3)	(-1,3)	(0,3)	(1,3)	(2,3)	(3,3)	(4,3)	(5,3)
(-5,2)	(-4,2)	(-3,2)	(-2,2)	(-1,2)	(0,2)	(1,2)	(2,2)	(3,2)	(4,2)	(5,2)
(-5,1)	(-4,1)	(-3,1)	(-2,1)	(-1,1)	(0,1)	(1,1)	(2,1)	(3,1)	(4,1)	(5,1)
(-5,0)	(-4,0)	(-3,0)	(-2,0)	(-1,0)	BS (0,0)	(1,0)	(2,0)	(3,0)	(4,0)	(5,0)
(-5,-1)	(-4,-1)	(-3,-1)	(-2,-1)	(-1,-1)	(0,-1)	(1,-1)	(2,-1)	(3,-1)	(4,-1)	(5,-1)
(-5,-2)	(-4,-2)	(-3,-2)	(-2,-2)	(-1,-2)	(0,-2)	(1,-2)	(2,-2)	(3,-2)	(4,-2)	(5,-2)
(-5,-3)	(-4,-3)	(-3,-3)	(-2,-3)	(-1,-3)	(0,-3)	(1,-3)	(2,-3)	(3,-3)	(4,-3)	(5,-3)
(-5,-4)	(-4,-4)	(-3,-4)	(-2,-4)	(-1,-4)	(0,-4)	(1,-4)	(2,-4)	(3,-4)	(4,-4)	(5,-4)
(-5,-5)	(-4,-5)	(-3,-5)	(-2,-5)	(-1,-5)	(0,-5)	(1,-5)	(2,-5)	(3,-5)	(4,-5)	(5,-5)

Fig. 2 cell identification

- **Communication range versus Communication cells:** By projecting the communication range to the cellular layer, hence considering the cells belonging to the communication range forming a square around certain cell as the **communication cells** of sensors belonging to that cell, as shown in figure 3.
- **Sensing range versus Sensing cells:** by projecting the sensing range of sensors belonging to a certain cell to the cellular layer the same way we did in the communication range, we can say that the **sensing cells** are the cells forming a square around that cell and belonging to the **sensing range** of the cell's sensors, as shown in figure 3.

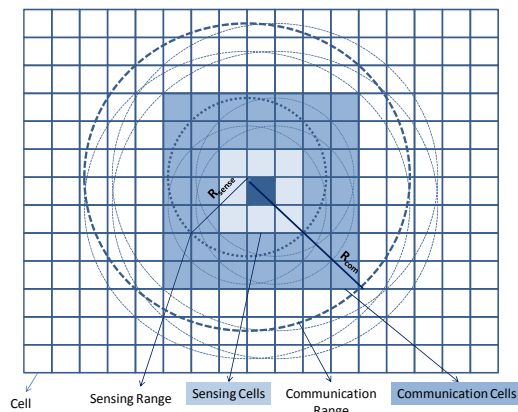


Figure 3 Communication cells and sensing cells

- **Processing Node:** In order to minimize the number of redundant packets transmitted to the base station, a processing node concept is introduced. The processing node of the cell is responsible for receiving all the packets sent to that cell and fuses these packets as a one aggregated packet to the next target cell. The processing node is elected for each cell as follows: The first sensor node locates itself in the cell claims itself as the processing node of that

cell. The processing node dissipates its energy faster than other nodes in the cell due to the aggregation and routing responsibilities assigned to it. When the processing node reaches a certain energy threshold  $E_{th}$  at the time instant  $T$ , it broadcasts a "request for a new processing node" message "Req-new-PN" to all nodes in the cell.  $E_{th}$  can be calculated as

$$E_{th} = E_{trans} + E_{active}$$

Where  $E_{trans}$  is the energy consumption of transmitting one message within the cell And  $E_{active}$  is the energy needed to keep a sensor active for one second.

Each node receives the "Req-new-PN" message broadcasts a message to all nodes in the cell to announce itself as the new processing node to that cell at time  $T_{announce}$ .  $T_{announce}$  is calculated as

$$T_{announce} = T + T_r$$

Where  $T_r$  is a random number between 0 and 1.

This guarantees that the new processing node will be announced in a time interval less than or equal to one second, hence before the old processing node dissipates all of its energy and dies. If the node receives one new processing node announcement before its  $T_{announce}$ , it aborts the process of claiming itself as the new processing node of the cell. A node detects a collision in two of the following cases: first if it receives two or more processing node announcements, in that case it will proceed in the process of claiming itself as the new processing node normally. The second case is when it sends an announcement itself and receives other announcements. In that case it will calculate a new  $T_{announce}$  and repeat the process of announcing itself as the new processing node all over again.

- **Communication Model:** Packets sent through the network can be categorized as "data packets", "Setup Packets", and "information packets". Data packets contain sensed data of a SN. SNs share data packets to form information packets. Setup packets are the packets that used within the cell of PN election. Information packets contain the location of the moving object and need to be fused to the Base station. Each packet, either data, setup or information packet, have a header contains the target cell id and the packet type. The target cell id is Null in the *data packet*, and is the same cell id as the sender's cell id in the setup packet. **Sending Process:** A SN sends a packet by broadcasting it thought the SN's communication range. **Receiving Process:** a SN receives a packet checks the target cell in the packet header. If the target cell is Null then the packet is a data packet and the SN stores the content of this packet to use it later in forming an information packet. On the other hand, if a SN receives a packet

with a non-Null target cell value, then SN checks the packet type, if it is a setup packet and belongs to the same cell as the receiver SN, it deals with it in the previously illustrated PN election process. The SN receives and fuses the information packet only if the SN is a PN and residing in the target cell, otherwise it just discards the information packet.

**The benefits of applying cellular layer to the network area are:**

- Cellular layer guarantees certain level of organization through the ad-hoc network structure.
- The routing between nodes in cellular arrangement is simple. The routing operations are applied to the cellular layer rather than nodes' layer. The node sends a packet through the route by fusing the packet to the next cell in the route through the network without necessity knowledge of the exact location of the next node in the route.
- The cellular architecture makes use of the inheritance property. In cellular architecture *scalability* of nodes is guaranteed as any node can be deployed to the network and locates itself to a certain cell, hence automatically inherits the cell properties in which it belongs, ex. The next routing cell, its parent cell if it is in a tree-based structure or its cluster head cell if it is in a cluster based structure, therefore operates within the network without an extra deployment costs.
- **Fault tolerance** of nodes is guaranteed. Node failures do not affect network operations, since sensing and routing operations are implemented in the term of cells rather than nodes.

**4. Energy aware Cellular architectures:**

The results showed that activating all the sensor nodes shorting the total network lifetime. Running an energy aware architecture prolongs the network lifetime compared to the worst case where all sensor nodes happen to be active. Therefore, the cellular architecture can make use of the *sensors' modes*. The sensors have different modes each consuming different amount of its battery power; these modes are defined in table 1.

Table1. Sensor node modes

Sensor node mode	Description
Stand-by	Sensing and communication components of the node are switched off and only the computation component is switched on.
Active	All sensor node three components, sensing , computation and communication, are switched on
Sleep	All sensor node three components, sensing , computation and communication are switched off

Energy aware cellular architecture is implemented by trying to put group of sensors in a standby mode in order to save their battery power thus maximizing the whole network life time and also reduce the information redundancy routed through the network. To guarantee uninterrupted routing, the processing node in each cell has to be in the active mode all the cell life time. The suggested Energy aware architectures are the Cellular Automata based architecture CA and the Radio Triggered based architecture RT.

#### 4.1. Cellular Automata Based Architecture:

Many areas of research have been benefited with the theories related to the cellular automata, mainly the areas that need to deal with systems that are constantly changing or have a random behavior as Wireless sensor networks. In [16][17], the authors introduced and used the cellular automata to simulate the large wireless sensor networks architecture. The suggested energy aware cellular automata based architecture applies the following rules:(1) A sensor that is in a standby mode on time  $t$  comes back to active mode on time  $t + 1$  if less than two sensors in the same cell that it belongs to are in active mode. (2) A sensor that is in active mode on time  $t$  go to standby modes on time  $t + 1$  if, on time  $t$ , two or more sensors in the same cell that it belongs to are in active mode.

The suggested cellular automata based architecture works as follows: In the beginning of simulation, all the sensor nodes deployed, are assumed to work in an active mode. The first sensor node in a cell locates itself is assumed to be the processing node PN of that cell. The PN of a cell does not perform the cellular automata rules and always works in an active mode. For each active sensor, not PN, after a random time  $T_r$  the active sensor reaches a checking point. In the checking point, the sensor node decides whether it stays in the active mode or goes to a standby mode by applying the previously declared cellular automata rules. Also the sensor node running in a standby mode goes to an active mode after a random time  $T_r$  and applies the cellular automata rules in order to decide the mode it will run under. In any checking point for any node in a cell, if the sensor finds zero active sensors, it assumes itself as the PN of that cell. After the entire energy of a sensor is consumed, the sensor goes to the sleep mode and becomes inactive; this process continues until all the sensor nodes' energy is consumed (i.e. all the cells are inactive). The responsibility of the PN in any cell is delegated through the active sensors at the time instance at which the PN sensor battery reaches the  $E_{th}$ , as described earlier. Cellular automata rules application packets sent through the cell are assumed to be setup packets.

#### 4.2. Radio Triggered Based Architecture:

In [18] radio-triggered power management, a special hardware component – a radio-triggered circuit – is

connected to one of the interrupt inputs of the processor. The circuit itself does not have any power supply. The node can enter a standby mode without periodic wake-up. When a power management message is sent by another node within a certain distance, the radio-triggered circuit collects enough energy to trigger the interrupt to wake up the network node. This is significantly different than using the radio transceiver to listen to messages because a listening radio transceiver requires help from the processor (or a radio sub-controller) to conduct channel monitoring and message parsing and the listening process consumes energy of the node. The radio-triggered circuit, on the contrary, is powered by the radio signals themselves. It is powered off when there are no suitable radio signals and naturally starts working when suitable radio signals arrive. Except for activating the wake-up interrupt, the radio-triggered circuit is independent of any other components on the node.

The suggested cellular radio triggered architecture works as follows: The first sensor locates itself in a cell is assumed to be the Processing node PN of that cell and works in an active mode. For any successive sensors locating themselves in that cell, they work in a standby mode. If the energy level of the current processing node reaches a certain threshold at time  $T$ , which means that this sensor node is running out of energy. Thus it sends a radio signal to the nodes in its cell to pick up the new PN. Each sensor node receives a radio signal waits a random time  $T_r$  between 0 and 1, and then goes to an active mode setting itself as the PN of that cell. If a standby sensor node received a radio signal at time  $T$  and picks up its random time  $T_r$ , then received another radio signal at time less than  $T+T_r$  then it aborts the operation of waking itself up to an active mode. This procedure guarantees that at any time instance only one sensor in a cell is in the active mode working as the processing node of that cell and all the other sensors in that cell are in a standby mode.

## 5. Locating moving object in the cellular architecture:

### 5.1. Sensing an object:

In [19] a survey of current available positioning techniques, their comparison, and subsequent recommendation of which technique is appropriate for use depending on specific application is presented. Using acoustic sensors, the trilateration method is the most appropriate method to use to locate an object. Whenever an object is moving in the network, it will be detected by a group of sensor nodes in its sensing range. At a  $t$  time instant, all the sensors that in the object's sensing range have to locate the object. Each sensor senses the object needs its measurement and two other measurements from other sensors to locate the object.



To apply the trilateration method to the suggested cellular architecture, active sensor nodes in the sensing cells of the object and not processing nodes sends a packet of its measurement locally inside the cell in which that sensor belongs and receive two other measurements to identify the objects position. If the sensor sensing an object is a Processing node sensor, then it broadcasts a packet of its measurement to a distance equals to two cells ahead. These Processing nodes broadcasts are made to two-cells ahead distance in order to ensure that in case all sensor nodes died in all sensing cells except for processing nodes, each processing node in a sensing cell will receive packets from other sensing cells and hence can locate the object. Then the object location is identified by its coordinates  $(X_{obj}, Y_{obj})$ .

Using the object's coordinates  $(X_{obj}, Y_{obj})$ , the cell in which the object resides can be identified and known as the **residential cell**. Only the sensor nodes belonging to the residential cell sensing cells are allowed to create information packets recording the object movement. Each information packet containing the location of the sensed object and the time instant at which the object is sensed at this specific location.

### 5.2. Routing packets to the base station:

After the information packets are created, the sensor nodes fuse them to the base station. The fusions are performed by the following scenarios:

#### 1. Direct transmission:

The radio transmission energy dissipation includes two parts of radio electronics energy and power amplifier energy. Generally the amplifier energy required for a successful transmission is much larger than the radio electronics energy and dominates the transmission energy dissipation.

#### • Sending packets Energy consumption simulation:

- To send a packet through the network, the energy consumption of the sending process depends on the distance between the sending and the receiving sensor nodes. Assume that the sending distance is defined as  $d$ ,  $d$  can be simulated as follows:
- If the packet is sent locally between two sensors inside one cell, then the maximum distance travelled by this packet will be

$$d = \text{Cell Diagonal} = d_0 / 4 = \sqrt{\frac{E_{fs}}{E_{mp}}} / 4$$

- If the packet is sent between two nodes in two different cells, defined as sending cell  $S$  and receiving cell  $R$ , then the maximum sending distance  $d$  between these two sensors can be calculated as:

$$d = (\text{CellDist}(S,R) + 1) * \text{Cell Diagonal}$$

- Where  $\text{CellDist}(S,R)$  is the number of cells between the sending cell  $S$  and the receiving cell  $R$ .  $\text{CellDist}(S,R)$  can be easily calculated assuming the sending cell is cell  $S(X_s, Y_s)$  and the receiving cell is  $R(X_r, Y_r)$

$$\text{CellDist}(S,R) = \text{Maximum} ( |X_s - X_r|, |Y_s - Y_r| )$$

- Deploying the distance estimation in the sending energy consumption equations allowing estimate for the sending process. The estimation is as follows: assume that sensor  $A$  residing in cell  $S$  is sending an  $m$ -bit packet to sensor  $B$  residing in cell  $R$ , and the  $\text{CellDist}(S,R) = \text{CellDist}$ , then the sending energy consumption can be calculated as:

$$E_{trans}(m, \text{CellDist}) = \begin{cases} m * (E_{elec} + \frac{E_{fs}^2 * (\text{CellDist} + 1)^2}{16 * E_{mp}}) & \text{where } \text{CellDist} \leq 3 \\ m * (E_{elec} + \frac{E_{fs}^2 * (\text{CellDist} + 1)^4}{256 * E_{mp}}) & \text{where } \text{CellDist} > 3 \end{cases}$$

#### 2. Multi-hop fusion:

According to the free space channel model, as described earlier, the minimum required amplifier energy is proportional to the square of the distance from the transmitter to the destined receiver ( $E_{trans} \propto d^2$ ) [19]. So the transmission energy consumption will increase greatly as the transmission distance rises. It means that sending packets directly from the sensor nodes to the base station consumes much energy than a multi-hop transmission. Therefore a multi-hop transmission routing is preferably applied in the cellular based architecture of our WSN. The multi-hop fusion will be applied as follows:

1. Each sensor senses a moving object and belongs to its sensing cells, creates a data packet to route.
2. If the sensor is not the processing node in the cell, the sensor sends its data packet to the processing node of the cell in which it belongs.
3. If the sensor is the processing node in the cell, it aggregates all the data packets sent to it by the sensors in the cell and creates one aggregated data packet to route.
4. The processing node determines the next **target cell** to which it will send its packet.
5. The **target cell** has to be at most three cells ahead towards the base station, which will be the farthest cell that can be reached in the communication cells of the cell in which the processing node belongs. The target cell can easily be identified by the numbering system defined above.
6. The processing node sends its aggregated packet to the processing node in the target

cell. The process are re implemented at the target cell to the next target cell until the packet reach the origin cell (0,0) in which the based station is located.

### 3. Static Tree-Based Routing :

The tree based routing depends on building a geographically designed tree to transmit data packets through it until it reaches the base station. The tree configuration is based upon the geographical location of the network's cells. Tree routing can reduce the data redundancy packets that can be transmitted to the base station where each parent cell will aggregate data packets created by its leaf cells.

#### • **Tree formulation**

1. Set the root cell for the tree as cell (0, 0).
2. All the cells belonging to the communication cells of the root cell are set as children cells for the root (0, 0). And these cells are considered as level 1 of the tree.
3. Determining the cells belonging to the next level in the tree. Those cells are in communication cells for the outer frame of current cell.
4. Each level corner cells are easily identified as the parent of the tree's next level.
5. Other cells belonging to outer frame of the next level are added to the parent cells in which distance between two parents should be at most 6 cells.
6. For each cell in the next level, assigning a parent cell that belonging to its communication cells and has no previously assigned parent.
7. Repeating steps 3:6 until all cells in the network area join the tree.

Therefore, all sensors belonging to a certain cell inherit their cell properties, either it is a parent cell and/or it is a child cell to a predefined parent. Insertion of new sensors in the network area needs no tree reformulation. The routing process in the tree based cellular architecture presented is implemented as follows:

1. The nodes that sense the moving object determine its residential cell.
2. Only the nodes belonging to the sensing cells of the residential cell of an object create data packets.
3. The data packets created are sent from the sensor node, which is not processing node, to the processing node of the cell in which it belongs.
4. The processing node aggregates the packets sent to it from the sensing sensors in the cell in which it belongs to one packet.

5. The processing node sends the aggregated packet to its assigned parent cell's processing node.
6. The processing node in the parent cell will aggregate data packets from its children cells and then fuses the aggregated packet to its own parent.
7. This procedure is repeated until the data packets delivered to the base station.

### 4. Aggregation-tree-Based Routing:

The geographical-tree based routing consumed more energy at parent cells, as parent cells are fixed cells in the network. This over-energy consumption at the parent cells causes them to die before the non-parent cells. Therefore, all the children of that died parent cell cannot transmit packets to the base station and hence leads to failure in that part of the network. In order to overcome this problem a dynamic aggregation tree for each sensed object is formulated. Therefore, at a time instant  $t$ , and target is moving through the network:

1. The nodes that sense the moving object determine its residential cell.
2. Only the nodes belonging to the sensing cells of the residential cell of an object create data packets.
3. The data packets created are sent from the sensor node, which is not processing node to the processing node of the cell in which it belongs.
4. The processing node aggregates the packets sent to it from the sensing sensors in the cell in which it belongs to one packet.
5. The Processing node defines the farthest cell in its communication cells towards the base station for the cell in which it belongs is its parent cell.
6. The aggregated packet is sent to the parent cell's processing node.
7. The processing node in the parent cell will aggregate data packets from all its children cells and then fuses the aggregated packet to its own parent, which is determined the same way in step 5.
8. This procedure is repeated until the data packets delivered to the base station.

### 5. Static-Cluster Based Routing:

In the conventional cluster architecture, clusters are formed statistically at the time of network deployment and the properties of each cluster are fixed such as number of members, area covered, etc. Static clustering has several drawbacks regardless of its simplicity, for example, static membership is not robust from fault-tolerance point of view and it prevents importing new sensor nodes in the network.

In the proposed architecture a geographically static clusters are formed through the network with a static

cell acting as a cluster head cell. In the cluster head cell each sensor node takes the responsibility as a cluster head node for the cluster successively. Simply the PN of the cluster head cell takes the responsibility as cluster head sensor.

The cluster is formed by the all the cells belonging to the communication cells of the cluster head cell. Consequently all the sensors belonging to those cells forming the cluster are considered as cluster member nodes. The cluster head cell is assumed to be located at the origin of the cluster and is reachable to all sensors belonging to the cluster by one message only. As declared previously, the maximum communication range that gives reasonable energy consumption in sending packets is three cells. So the clusters are assumed to be of size  $7 \times 7$  cells, having the cluster head node at the origin of the cluster.

The proposed clustering technique combines the simplicity of the static clustering, as no need for cluster periodic reformation. It also can have the properties of dynamic clustering as it can handle nodes scalability of the network. If a new sensor node imported in the network, it automatically locates itself and determines its cell, consequently determining to which cluster it belongs without a need for a joining messages between the imported node and the cluster head node. The cluster routing is implemented as follows:

1. The nodes that sense the moving object determine its residential cell.
2. Only the nodes belonging to the sensing cells of the residential cell of an object create data packets.
3. The data packets created are sent from the sensor node, which is not processing node to the processing node of the cell in which it belongs.
4. The processing node aggregates the packets sent to it from the sensing sensors in the cell in which it belongs to one packet.
5. The processing node sends the aggregated packet to the cluster head node in the cluster head cell for the cluster in which it belongs.
6. The cluster head node aggregates all data packets from its member cells and then fuses the aggregated packet to the base station in a multi-hop transmission fashion through the elected processing nodes in the route cells, as described earlier.
6. Dynamic Cluster Routing:

In contrast, dynamic clustering offers several advantages where clusters are formed dynamically depending on occurrence of certain events. For the suggested dynamic cluster based routing, the cluster is formed through the tracking process based on the geographic position of the object being tracked. Since sensors don't statistically form a cluster, they may belong to different clusters at different timings. As

only one cluster is active at a time, redundant data and interference is reduced.

Forming the cluster in the dynamic clustering is as follows: The residential cell of an object is assumed to be the current cluster head cell. The processing node in the current cluster head cell will be the current cluster head node. The cluster member cells will be the cells in the sensing cells of the residential cell, which are one cell ahead surrounding the cluster head cell. Consequently, all the nodes belonging to the cluster member cells are assumed as member nodes of the current cluster.

The object tracking process will be performed as follows:

1. The nodes that sense the moving object determine its residential cell.
2. Only the nodes belonging to the sensing cells of the residential cell of an object create data packets.
3. The data packets created are sent from the sensor nodes, which is not processing node to the processing node of the cell in which they belong.
4. The processing node aggregates the packets sent to it from the sensing sensors in the cell in which it belongs to one packet.
5. The aggregated data packets created at the processing nodes in the member cells are sent to the cluster head node (which is the processing node of the residential cell).
6. The cluster head node aggregates data from all current cluster member cells and fuses the aggregated packet containing the object location and time instant using the previously described multi-hop aggregated routing protocol through the processing nodes of the route to the base station.

The advantage of using the dynamic cluster head routing through the cellular architecture is that, whenever an object moves the cluster is reformed by the cluster head cell, which will always be the residential cell of the object. Also the member cells of the cluster will always be the cells surrounding the cluster head cell. Therefore, the cluster head node will also be automatically recognized by all the cluster member nodes as the processing node of the cluster head cell. No extra messages need to be sent to determine the cluster size, member nodes, or the cluster head node for the current cluster. That means no overhead energy consumption is needed for forming the dynamic cluster while the object is moving.

## 6. Simulation model

The proposed architecture is implemented to simulate a large stochastic and densely deployed wireless sensor network. The implementation was constructed as follows:

- **Network modeling:** The coverage area consists of 33\*33 squared cells each containing 9 sensor nodes, with a random waypoint moving object with constant velocity crossing one cell per second .
- **Energy Modeling:** WSN sensor energy at the beginning of operation = 2 J . The energy of an WSN sensor node in *active mode* decreases by 0.0165 J every time step (in our case every 1 sec). The energy of a WSN sensor node in *stand-by mode* decreases by 0.00006 J every time step (in our case every 1 sec). Other parameters  $E_{fs} = 10 \text{ pJ/bit/m}^2$ ,  $E_{mp} = 0.0013 \text{ pJ/bit/m}^4$ ,  $E_{elec} = 50\text{nJ/bit}$ ,  $m = 512 \text{ byte}$ .
- **Performance measurements:**
  - Global energy of the network: the sum of the residual energy remaining on all nodes in a specific point in time.
  - Network lifetime: the time elapsed from the start of the simulation until all nodes run out of energy.
  - Sensing energy consumption
  - Overhead communication energy consumption which is caused by transmitting all the non information packets, and this takes place in the process of PN election and cellular Automata rules application.
  - Routing-Protocols energy consumption comparison.
  - Redundancy reduction which is the sensing to receiving packets ratio.

• **Simulation results:**

Each point on the curve is the average of 20 runs with 90% confidence interval. Figure 4 shows the reduction in energy consumption caused by using the multi-hop routing instead of direct transmission of packets sensing a moving object in an AA network. The multi-hop routing only uses maximum of 1.5% of the energy used when transmitting packets directly to the base station.

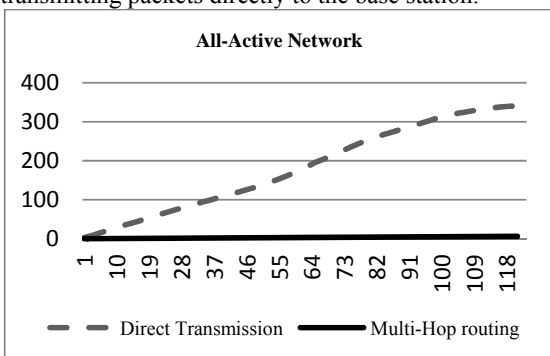


Fig. 4 Direct transmission versus multi-hop routing in AA network

Figure 5 shows a comparison of the various types of routing protocols used to transmit packets recording a moving object in an AA network. Results show that the aggregate tree routing protocol uses a minimum of 87.9% of the energy used by the multi-hop routing protocol. The static cluster routing protocol uses minimum of 74.5% of the energy used by the multi-hop routing protocol. The static tree based routing

protocol uses a minimum of 71.6% of the energy used by the multi-hop routing protocol. Finally the best reduction given by the dynamic cluster based routing protocol, with minimum of only 67.7% of the energy used by the multi-hop routing protocol.

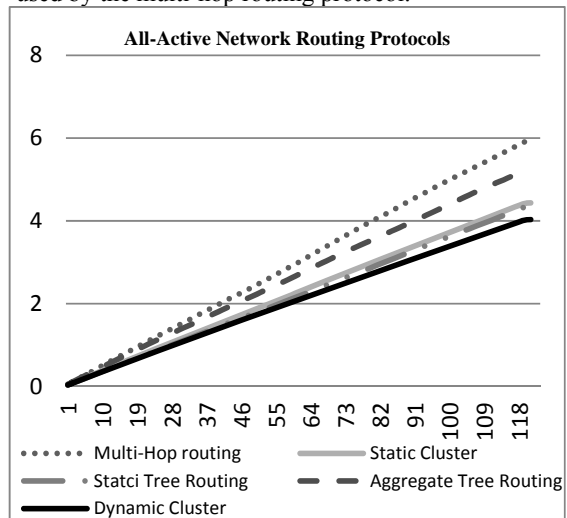


Fig 5 comparison of routing protocols over AA network

Figure 6 shows the number of delivered packets for all the routing protocols in the AA network. The results show that the redundancy reduction is a minimum in the dynamic cluster based routing, with only 6% of the sensed packets received at the base station. However only 7.4% of the sensed packets delivered to the base station with the tree based routing protocol, 10% with the static cluster based routing protocol, 30.7% with the aggregate tree routing protocol and 54.8% with the multi-hop routing protocol.

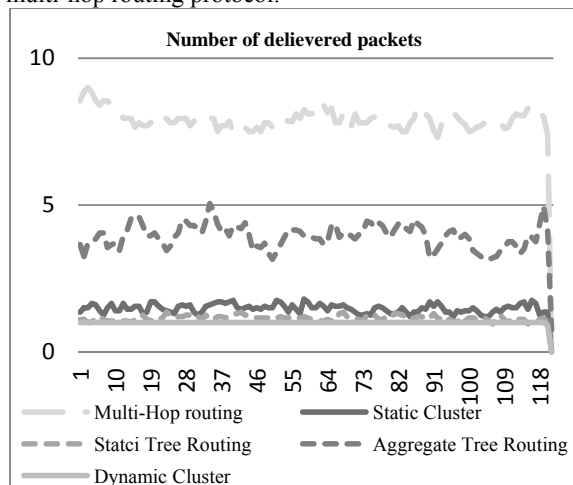


Fig 6 redundancy reductions in packets delivery in AA network

Figure 7 shows the reduction in energy consumption using the multi-hop routing instead of direct transmission of packets sensing a moving object in a CA network. The multi-hop routing only uses maximum of 2.2% of the energy used when transmitting packets directly to the base station.

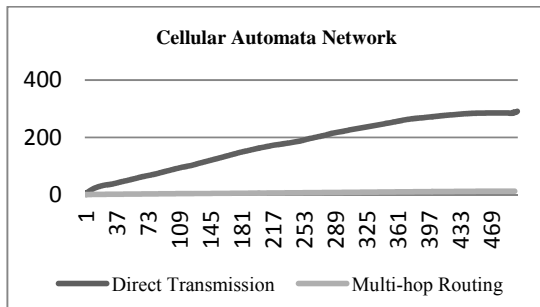


Fig. 7 Direct transmission versus multi-hop routing in CA network

Figure 8 shows a comparison of routing protocols used to transmit packets recording a moving object in a CA network. Results show that the aggregate tree routing protocol uses a minimum of 78.2% of the energy used by the multi-hop routing protocol. The static cluster routing protocol uses minimum of 48.3% of the energy used by the multi-hop routing protocol. The static tree based routing protocol uses a minimum of 46.3% of the energy used by the multi-hop routing protocol. Finally the best reduction given by the dynamic cluster based routing protocol, with minimum of only 38.1% of the energy used by the multi-hop routing protocol.

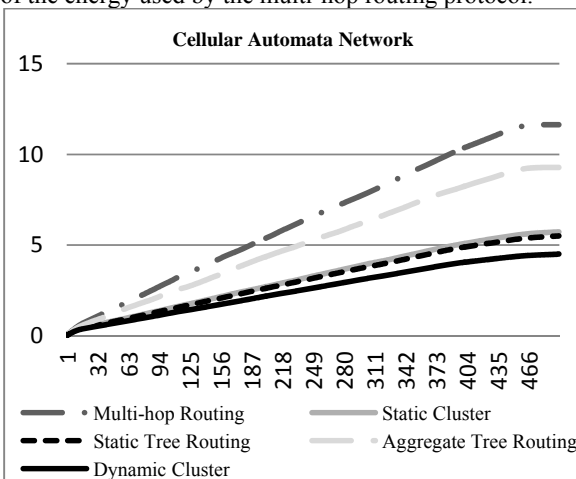


Fig. 8 comparison of routing protocols over CA network

Figure 9 shows the number of delivered packets ratio for all the routing protocols in the CA network. The results show that the redundancy reduction is maximized in the dynamic cluster based routing, then tree based routing protocol, Static cluster based routing protocol, aggregate tree routing protocol and finally the multi-hop routing protocol.

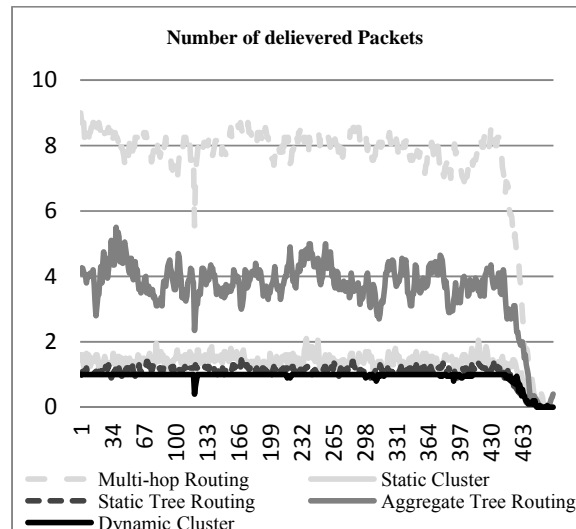


Fig. 9 redundancy reductions in packets delivery in CA network

Figure 10 shows the reduction in energy consumption using the multi-hop routing instead of direct transmission of packets sensing a moving object in a RT network. The multi-hop routing only uses maximum of 5.1% of the energy used when transmitting packets directly to the base station.

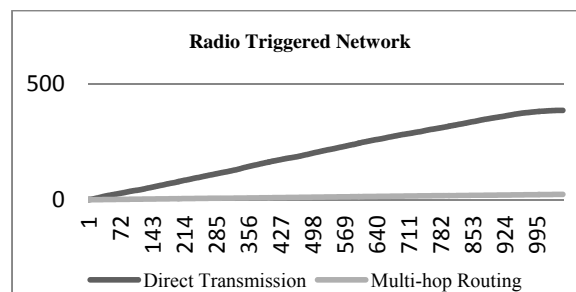


Fig. 10 Direct transmission versus multi-hop routing in RT network

Figure 11 shows a comparison of routing protocols used to transmit packets recording a moving object in a RT network. Results show that the aggregate tree routing protocol uses a minimum of 73.5% of the energy used by the multi-hop routing protocol. The static cluster routing protocol uses minimum of 39.2% of the energy used by the multi-hop routing protocol. The static tree based routing protocol uses a minimum of 37.1% of the energy used by the multi-hop routing protocol. Finally the best reduction given by the dynamic cluster based routing protocol, with minimum of only 28.1% of the energy used by the multi-hop routing protocol.

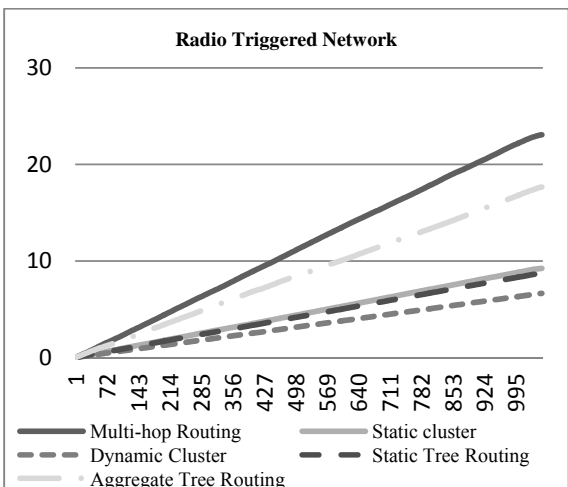


Fig. 11 comparison of routing protocols over RT network

Figure 12 shows the Sensed to delivered packets ratio for all the routing protocols in the RT network. The results show that the redundancy reduction is maximized in the dynamic cluster based routing, then tree based routing protocol, Static cluster based routing protocol, aggregate tree routing protocol and finally the multi-hop routing protocol.

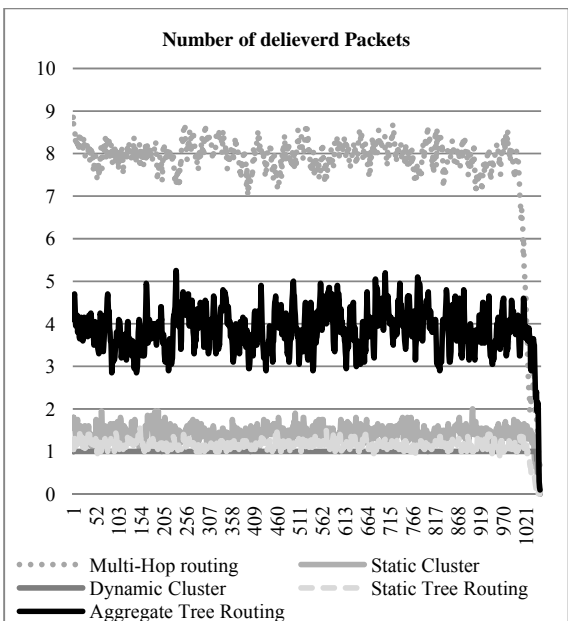


Figure 12 redundancy reductions in packets delivery in RT network

Next the figures 13:21 compare between the three proposed networks. Figure 13 shows the network lifetime of the AA, CA, and RT networks. The results show that the CA network lifetime is around 4 times the AA network and the RT network is over 8 times the AA network.

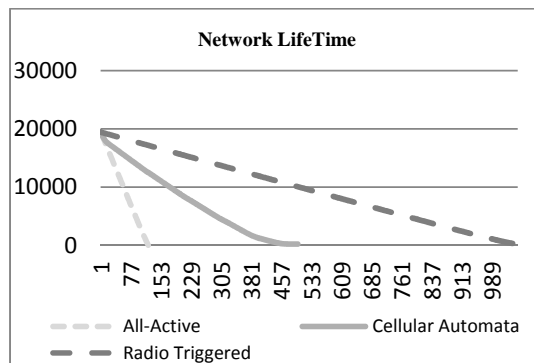


Fig. 13 the AA, CA, and RT networks lifetime

Figure 14 shows the Energy consumed by the sensing process in the three networks. The results show that the CA network consumes a minimum of 29.4% of the energy consumed in sensing a moving object in the AA network. Also the results show that the RT network only consumes at least 12.8% of the AA network sensing energy.

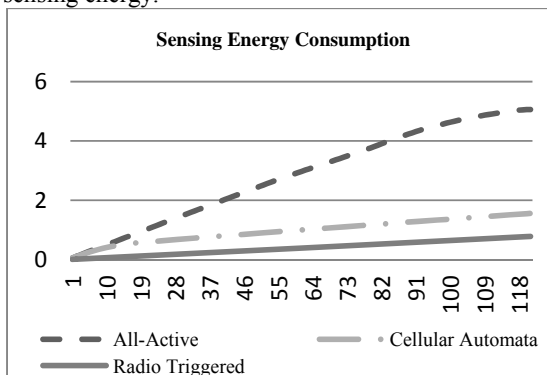


Fig. 14 the AA, CA, and RT Sensing

Figure 15 shows energy consumed by overhead communication for non-object location related messages. Results give less consumption at RT network.

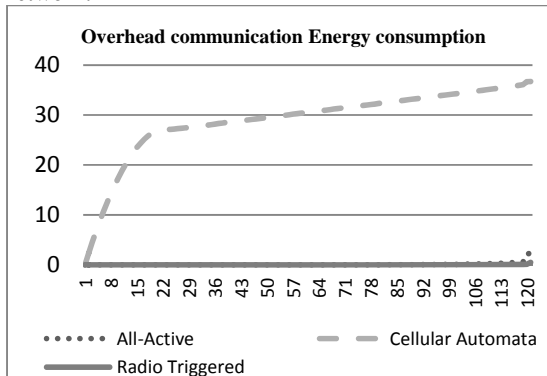


Fig. 15 the AA, CA and RT overhead communication

Figure 16 shows the Energy consumed by transmitting packets directly in the three networks recording an object's movements. The results show that the CA network consumes a minimum of 27.5% of the energy consumed in the AA network. Also the results show that the RT network only consumes 8.5% of the energy consumed in the AA network.

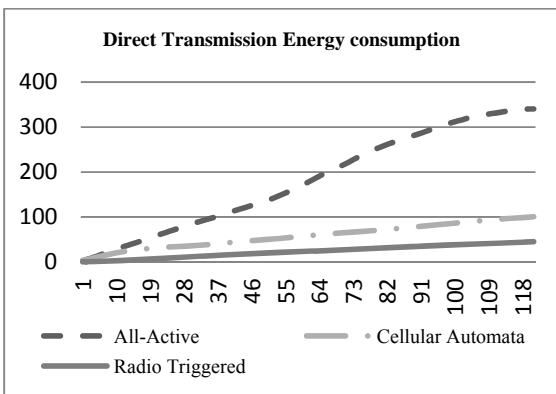


Fig. 16 AA, CA, and RT Direct Transmission Energy consumption  
 Figure 17 shows the Energy consumed by Multi-hop routing protocol application in the three networks recording an object's movements. The results show that the CA network consumes a minimum of 75.4% of the energy consumed in the AA network. Also the results show that the RT network only consumes 38.5% of the energy consumed in the AA network.

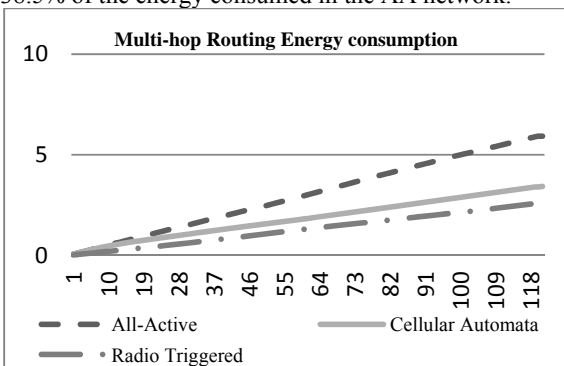


Fig. 17 AA, CA, and RT Multi-hop routing Energy consumption  
 Figure 18 shows the Energy consumed Static cluster based routing application in the three networks recording an object's movements. The results show that the CA network consumes a minimum of 38.5% of the energy consumed in the AA network. Also the results show that the RT network only consumes 23% of the energy consumed in the AA network.

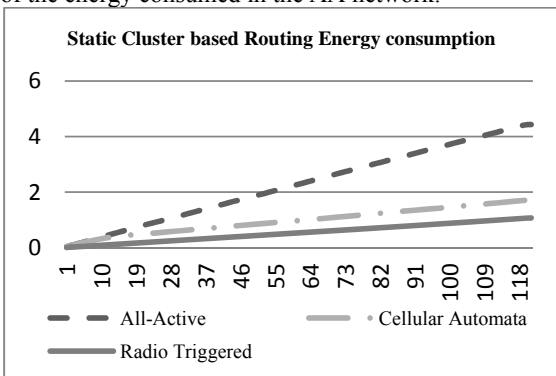


Fig. 18 AA, CA, and RT static cluster based routing Energy consumption  
 Figure 19 shows the Energy consumed Static Tree based routing application in the three networks recording an object's movements. The results show

that the CA network consumes a minimum of 38% of the energy consumed in the AA network. Also the results show that the RT network only consumes 21.8% of the energy consumed in the AA network.

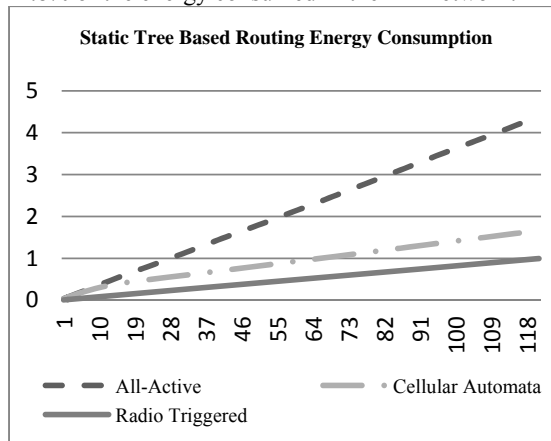


Fig. 19 the AA, CA, and RT Static Tree based routing Energy consumption

Figure 20 shows the Energy consumed Aggregate Tree based routing application in the three networks recording an object's movements. The results show that the CA network consumes a minimum of 50.9% of the energy consumed in the AA network. Also the results show that the RT network only consumes 28.7% of the energy consumed in the AA network.

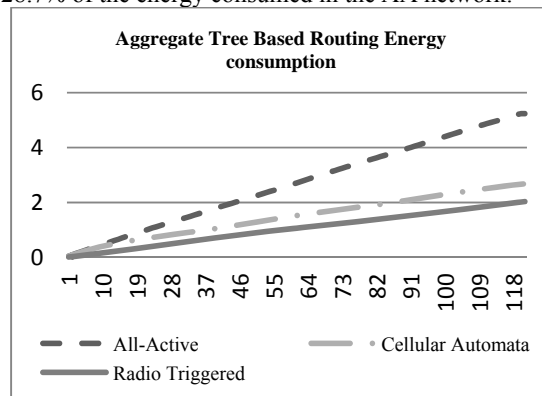


Fig. 20 the AA, CA, and RT Aggregate tree based routing Energy consumption

Figure 21 shows the Energy consumed Dynamic cluster based routing application in the three networks recording an object's movements. The results show that the CA network consumes a minimum of 34.6% of the energy consumed in the AA network. Also the results show that the RT network only consumes 17.1% of the energy consumed in the AA network.

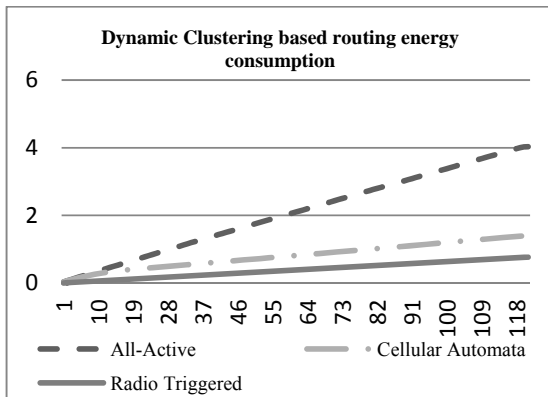


Figure 21 the AA, CA, and RT dynamic cluster based routing Energy consumption

## 7. Conclusion

In this paper, we have presented an approach to manage large wireless sensor networks based on cellular architecture. This approach allows us to create scenarios with a huge number of sensor nodes. It also provides scalability in order to insert new nodes in the network during its execution time with minimal overhead energy consumption due to the inheritance feature provided to each sensor joining the cellular network. The architecture proposed also allows fault tolerance in case of nodes failures for any abnormal or environmental situations. Data communication through the cellular architecture has been addressed through locating moving object application.

Two energy aware solutions have been applied to the proposed cellular architecture, organizing sensors operational states, in order to prolong the network lifetime. The suggested solutions also reduce the redundancy and communication overhead through the network. The first solution applies cellular automata rules, CA network, to the cellular architecture. And the second solution depends on a radio triggered sensor nodes to be applied to the network area, RT network.

Tracking Moving object application has been suggested to test the cellular architectures performance. Routing algorithms has been specially designed to work through the cellular architecture and benefits of the inheritance feature that can be used in the network cells. All Multi-hop, Static Tree, Aggregated Tree, Static cluster, and dynamic cluster based routing protocols have been designed.

A simulator has been developed to measure the performance of the suggested architectures. The obtained results compare between the three proposed architectures for overhead communication, sensing and routing protocols for tracking moving object through the network.

The results indicate that the radio triggered proposed architecture gives the longest network life time with less communication and less overhead cost over the all active and the cellular automata architectures.

Also a comparison between all the routing algorithms has been performed indicating a lowest routing energy consumption with best redundancy reduction for the dynamic cluster based routing protocol.

## References

- [1] W. Ye, J. Heidemann, D. Estrin, "An Energy-Efficient MAC Protocol for Wireless Sensor Networks", Proceedings of INFOCOM 2002, New York, June 2002, pp.1567-1576.
- [2] Sekine M., Nakamura S., Sezaki K., "An Energy-Efficient Protocol for Active/Sleep Schedule Synchronization in Wireless Sensor Networks", Asia-Pacific Conference on Communications (APCC), August 2006, pp.1-5.
- [3] F. Akyildiz, W. Su, Y. Sankarasubramaniam, E. Cayirci, "A survey on sensor networks", IEEE Communications Magazine, August 2002, Volume 40, Issue 8, pp.102-114.
- [4] J. L. Hill, D. E. Culler, "Mica: a wireless platform for deeply embedded networks", IEEE Micro, November-December 2002, Volume 22, Issue 6, pp.12-24.
- [5] Brain, M., How Motes Work, <http://computer.howstuffworks.com/mote.htm/printable>, Retrieved on August 16, 2008.
- [6] Carlos E. Otero, Ivica Kostanic, Luis D. Otero, "Development of a Simulator for Stochastic Deployment of Wireless Sensor Networks" JOURNAL OF NETWORKS, VOL. 4, NO. 8, OCTOBER 2009.
- [7] Bhatti, S., Xu, J., "Survey of target tracking protocols using wireless sensor network", In: Proc. of ICWMC. P. 110-115, 2009.
- [8] W.-P. Chen, J.C. Hou and L. Sha, "Dynamic Clustering for Acoustic Target Tracking in Wireless Sensor Networks," in Proceeding of 11th IEEE International Conference on Network Protocols (ICNP'03), Atlanta, Georgia, USA, pp. 284-294, November 2003.
- [9] S. C. Chuang, "Survey on Target Tracking in Wireless Sensor Networks." Dept. of Computer Science - National Tsing Hua University, (5/26/2005).
- [10] Huang, Q., Lu, C., and Roman, G.-C., "Spatiotemporal Multicast in Sensor Networks," Proceedings of the First, International Conference on Embedded Networked Sensor Systems (SenSys 2003), November 2003, pp. 205-217.
- [11] Rapaka and S. Madria, "Two Energy Efficient Algorithms For Tracking Objects In A Sensor Network," Wireless Communication Mobile Computing, 2006.
- [12] Santi, P., "Topology Control in Wireless Ad Hoc and Sensor Networks," ACM Computing Surveys, 2005.
- [13] Karl, H., Willig, A., Protocols and Architectures for Wireless Sensor Networks, Wiley, 1st Edition, 2007.
- [14] L. Doherty, K. S. J. Pister, and L. E. Ghaoui, "Convex position estimation in wireless sensor networks," in Proc. IEEE Infocom 2001, vol. 3, (Anchorage AK), pp. 1655-1663, Apr. 2001.
- [15] S. Meguerdichian, F. Koushanfar, M. Potkonjak, M. Srivastava, "Coverage Problems in Wireless Ad Hoc Sensor Networks," IEEE INFOCOM'01, vol 3, pp. 1380-1387, 2001.
- [16] R. O. Cunha, A. P. Silva, A. A. F. Loureiro and L. B. Ruiz, "Simulating large wireless sensor networks using cellular automata," in ANSS '05: Proceedings of the 38th Annual Symposium on Simulation, pp. 323-330, 2005.
- [17] Qela, B., Wainer, G., Mouftah, H., "Simulation of Large Wireless Sensor Networks Using Cell-DEVS," Proceedings of the Winter Simulation Conference on networking and communication, pp. 3189-3200, 2009.
- [18] Lin Gu, John A. Stankovic, Radio-Triggered Wake-Up for Wireless Sensor Networks, Real-Time Systems, v.29 n.2-3, p.157-182, March 2005.
- [19] T. Rappaport, Wireless Communications: Principles & Practice. Englewood Cliffs, NJ: Prentice-Hall, 1996.



# Security Analysis of Routing Protocols in Wireless Sensor Networks

Mohammad Sadeghi<sup>1</sup>, Farshad Khosravi<sup>2</sup>, Kayvan Atefi<sup>3</sup>, Mehdi Barati<sup>4</sup>

<sup>1</sup>Faculty of Computer and Mathematical Sciences, UiTM,  
Shah Alam, Malaysia

<sup>2</sup>Faculty of Electrical Engineering, Islamic Azad University  
Eslam Abad Gharb,Iran

<sup>3</sup>Faculty of Computer and Mathematical Sciences, UiTM,  
Shah Alam, Malaysia

<sup>4</sup>Faculty of Computer Science, UPM, Malaysia

## Abstract

In this paper, I describe briefly some of the different types of attacks on wireless sensor networks such as Sybil, HELLO, Wormhole and Sinkhole attacks. Then I describe security analysis of some major routing protocols in wireless sensor network such as Directed Diffusion, TinyOS beaconing, geographic and Rumor routings in term of attacks and security goals. As a result I explain some secure routing protocols for wireless sensor network and is discussed briefly some methods and policy of these protocols to meet their security requirements. At last some simulation results of these protocols that have been done by their designer are mentioned.

**Keywords:** Security Analysis, Routing Protocols, Wireless Sensor Networks (WSN), attacks in network

## I. INTRODUCTION

Wireless sensor networks (WSNs) have gained worldwide interest in these years. Advances in Microelectronic Systems and low power radio technologies have created low-cost, low-power, multi-functional sensors devices, which can sense, measure, and collect information from the environment and transmit the sensed data to the user by a radio transceiver. Sensor nodes can use battery as a main power source and harvest power from the environment like solar panels as a secondary power supply.

An unstructured WSN is a network that contains a dense collection of sensor nodes and can be automatically organized to form an ad hoc multihop network to communicate with each other. On the other hand, a structured WSN deploys all or some of the sensor nodes in a pre-planned manner. So, it has a lower network maintenance and management cost.

WSNs can be used for many applications such as military target tracking and surveillance, natural

disaster relief, biomedical health monitoring, environment exploration and agricultural industry [5]. The architecture of commonly used WSN is as depicted in figure 1.

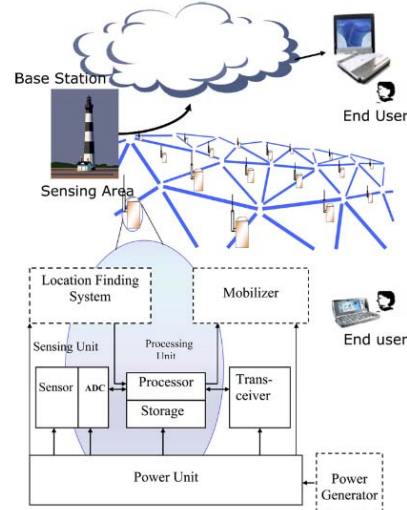


Fig.1 The architecture of commonly used in WSN

Wireless sensor networks like any wireless technology are susceptible to several security attacks due to the broadcast nature of transmission medium [1]. There are constraints in incorporating security into a WSN such as limitations in storage, communication, computation, and processing capabilities. To Design a security protocol we need to understand these limitations and achieve acceptable performance with security measures to meet the needs of an application.

In this paper, I describe briefly some of the different types of attacks on wireless sensor networks and also security analysis of some major routing protocols in wireless sensor network in term of design and security goals.

## II. THREAT MODEL

Attacks on wireless sensor network can be classified to mote-class attacks and laptop-class attacks. In the mote-class attacks, the attacker has access to a few sensor nodes with similar capabilities. On the other hands, a laptop-class attacker may have access to more powerful devices, like laptops or their equivalent. They may have greater battery power, a more capable CPU, a high-power radio transmitter, or a sensitive antenna and can do more than an attacker with only ordinary sensor nodes [3].

Another classification in attacks on wireless sensor network is based on the outsider or insider attacks. In insider attack a compromised node was captured by an adversary and may possess all the secret keys and be capable of participating in the communications and disrupting the network. In contrast, outsider attacks, where the attacker has no special access to the sensor network. The outsider attacks are achieved by unauthorized nodes that can easily eavesdrop on the packets exchanged between sensor nodes due to the shared wireless medium [2].

Based on the network layers, [6] cites another classification of attacks on wireless sensor network. Attacks at physical layer: Jamming is one of the most important attacks at physical layer. Aiming at interfering with normal operations, an attacker may continuously transmit radio signals on a wireless channel. An attacker can send high-energy signals in order to effectively block wireless medium and to prevent sensor nodes from communicating. This can lead to Denial-of-Service (DoS) attacks at the physical layers.

Attacks at link layer: The functionality of link layer protocols is to coordinate neighboring nodes to access shared wireless channels and to provide link abstraction to upper layers. Attackers can deliberately violate predefined protocol behaviors at link layer. For example, attackers may induce collisions by disrupting a packet, cause exhaustion of nodes' battery by repeated retransmissions, or cause unfairness by abusing a cooperative MAC layer priority scheme. All these can lead to DoS attacks at the link layers.

Attacks at network layer: In WSNs, attacks at routing layer may take many forms. This kind of attacks will be discussed in following.

Attacks targeting at WSN services and applications: basically, to prevent this kind of attack localization and aggregation are used.

## III. ATTACKS ON ROUTING PROTOCOLS IN WIRELESS SENSOR NETWORKS

Some of network layer attacks on wireless sensor networks are listed as follow:

### A) *Eavesdropping:*

Since transport medium in wireless sensor network use broadcasting feature, so any adversary with a strong receiver could eavesdrop and intercept transmitted data. Information like location of node, Message IDs, Node IDs, timestamps, application specific information can be retrieve by an intruder. To prevent these problems we should use strong encryption techniques [1].

### B) *Denial of service:*

In a Denial-of-Service (DoS) attack, an adversary attempts to disrupt, corrupt or destroy a network. It reduces or eliminates a network's capacity to perform its expected function [2].

### C) *Message tampering:*

Malicious nodes can tamper with the received messages thereby altering the information to be forwarded to the destination. At the destination side, the Cyclic Redundancy Code (CRC) would be computed. The redundancy check fails and it would result in dropping the packet. If the CRC check was successful then the destination node would accept wrong information [2].

By spoofing or altering or replaying routed information, false messages can be generated, routing loops can be created, latency of the network can be increased, etc. The motivation for mounting a replay attack is to encroach on the authenticity of the communication in WSNs [7].

### D) *Selective forwarding:*

In a selective forwarding attack, malicious nodes may refuse to forward certain messages and simply drop them, ensuring that they are not propagated any further. A simple form of this attack is when a malicious node behaves like a black hole and refuses to forward every packet she sees. By this, neighboring nodes will conclude that she has failed and decide to seek another route. A more subtle form of this attack is when an adversary selectively forwards packets. An adversary interested in suppressing packets originating from a select few nodes can reliably forward the remaining traffic and limit suspicion of her wrongdoing.

Selective forwarding attacks are typically most effective when the attacker is explicitly included on the path of a data flow. However, it is conceivable an adversary overhearing a flow passing through neighboring nodes might be able to emulate selective forwarding by jamming or causing a collision on each forwarded packet of interest [3].

**E) Sinkhole attacks:**

In a sinkhole attack, the adversary manipulates the neighbouring nodes to attract nearly all the traffic from a particular area through a compromised node and create a sink as shown in figure 2. This malicious sink can now not only tamper with the transmitted data but can also drop some vital data and lead to other attacks like eavesdropping and selective forwarding. Sinkhole attacks usually make a compromised node that is more attractive to neighbouring sensor nodes than the routing algorithm. This could be approached by spoofing or replaying an advertisement for an extremely high quality route to a sink. Therefore, all the surrounding node of the adversary will start forwarding packets destined for a sink through the adversary, and also propagate the attractiveness of the route to their neighbours [2].

[3] Noted that the reason sensor networks are particularly susceptible to sinkhole attacks is due to their specialized communication pattern. Since all packets share the same ultimate destination, a compromised node needs only to provide a single high quality route to the base station in order to influence a potentially large number of nodes.

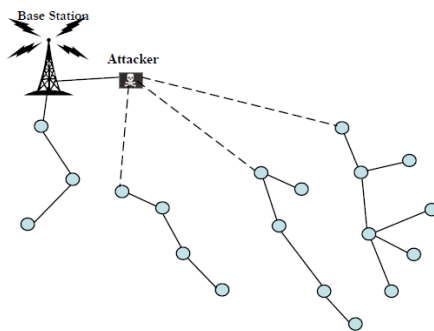


Fig.2: Sinkhole attacks

**F) Wormhole attacks:**

In this kind of attack, an adversary receives messages by making a tunnel and a low-latency link in one part of the network and replays them in a different part as shown in figure 3. An adversary could convince nodes who would normally be multiple hops from a sink that they are only one or two hops away via the wormhole. This would not only make some confusion in the routing mechanisms but would also create a sinkhole since the adversary on the other side of the wormhole can pretend to have a high quality route to the sink, potentially drawing all traffic in the surrounding area. An adversary that is situated near the sink may be able to completely disrupt routing by creating a well-placed wormhole [2].

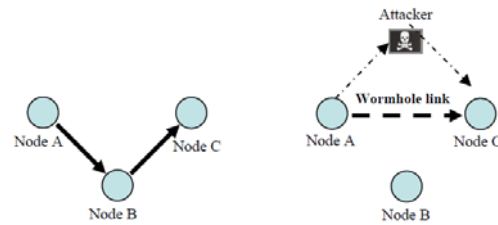


Fig.3: Normal Network (left), Wormhole Attack (right)

**G) Sybil attacks:**

In a Sybil attack, a single malicious node illegitimately presents multiple identities to other nodes in the network. The Sybil attack can significantly decrease the effectiveness of fault-tolerant schemes such as distributed storage, disparity and multipath routing, and topology maintenance. The Sybil attacks can take advantage of different layers to make service disruption. This attack at the routing layer will help the malicious node to draw in large amounts of network traffic to go through the same entity. This creates a sinkhole and as a result the attacker can do selective forwarding on received data [2].

In addition to defeating distributed data storage systems, the Sybil attack is also effective against routing algorithms, data aggregation, voting, fair resource allocation and foiling misbehavior detection. Regardless of the target all of the techniques involve utilizing multiple identities. For example, in a sensor network voting scheme, the Sybil attack might utilize multiple identities to generate additional “votes” [4].

**H) HELLO Attack:**

Nodes in WSNs learn about their neighboring nodes through HELLO packets. Every node advertises its presence to neighboring nodes by broadcasting HELLO packets. In HELLO attack, a malicious node follows the same technique. It uses transmission power high enough to reach the nodes that are very far away from its physical location which convinces the receivers of its advertised packets that it is a legitimate neighboring node as shown in Figure 4. Generally routing protocols of WSN depend on localized exchange of routing information to maintain routing topology and flow control [3].

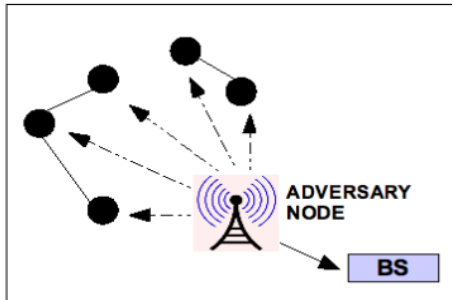


Fig. 4: HELLO attack

**1) Acknowledgement spoofing:**

Several sensor network routing algorithms rely on implicit or explicit link layer acknowledgements. An adversary can spoof link layer acknowledgment for “overheard” packets addressed to neighboring nodes to convince the sender that a weak link is strong or that a dead or disabled node is alive. By this attack a routing protocol may select the next hop in a path using link reliability [3].

**IV. ANALYSIS OF ROUTING PROTOCOLS IN WIRELESS SENSOR NETWORKS IN TERM OF ATTACKS AND COUNTERMEASURES**

All of the proposed sensor network routing protocols are highly susceptible to attack [3]. Some important routing protocols and relevant attacks will be discussed in following.

**A. Directed diffusion**

As [7] cites, Directed Diffusion is a data centric protocol for drawing information out of a sensor network. The base station asks for data by broadcasting interests. An interest is a task request that needs to be done by the network. Among the route, nodes keep propagating the interests until the nodes that can satisfy the interests are reached. Each node that receives the interests sets up a gradient toward the origin node. A gradient contains an attribute value and direction. As shown in Figure 5 when node B receives an interest from node A, it includes  $A(\Delta)$  in its gradient. When node C receives an interest from node A through node B, it includes  $B(2\Delta)$  in its gradient. On the other hand, when node C receives an interest from node A, it includes  $A(\Delta)$  in its gradient. When the data matches the interest (event), path of information, flows to the base station at low data rate. Then the base station recursively reinforces one or more neighbors to reply at a higher data rate. Alternatively, paths may be negatively reinforced as well.

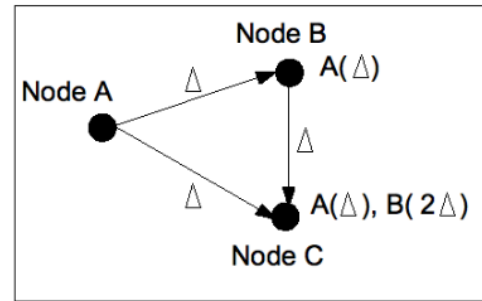


Fig. 5: Gradient set up in Directed Diffusion Routing Protocol

There is a multipath variant of directed diffusion as well. After the primary dataflow is established using positive reinforcements, alternate routes are recursively established with maximal disjointedness by attempting to reinforce neighbors not on the primary path [8].

It becomes an easy task for the attacker to eavesdrop the interest in this protocol. After an adversary receives an interest flooded from a legitimate base station, it can simply replay that interest with herself listed as a base station. When the response for that interest is sent, apart from the base station, the adversary would also be receiving them [7], [3].

When sources begin to generate data events, an adversary node might attack a data flow and cause to flow suppression. It is an instance of denial-of-service attack. The easiest way to suppress a flow is to spoof negative and positive reinforcements. It can also influence the path taken by a data flow. For instance, after receiving and rebroadcasting an interest, an adversary interested in directing the resulting flow of events through herself would strongly reinforce the nodes to which the interest was sent while spoofing high rate, low latency events to the nodes from which the interest was received. By using the above attack to insert herself onto the path taken by a flow of events, an adversary can gain full control of the flow. She can modify and selectively forward packets of her choosing [3].

On the other hand a laptop-class adversary can exert greater influence on the topology by creating a wormhole between one node that located next a base station and other node located close to where events are likely to be generated. Interests advertised by the base station are sent through the wormhole [7].

[3] Shows that the combination of the positive and negative reinforcements pushes data flows away from the base station and towards the resulting sinkhole.

A single adversary can use the Sybil attack against her neighbors even in the multipath version. A neighbor will be convinced it is maximizing diversity by reinforcing its next most preferred neighbor not on the primary flow when in fact this neighbor is an alternate identity of adversary [3].

### B. TinyOS beaconing

This protocol builds a spanning tree with a base station as the parent for all the nodes in the network. Periodically the base station broadcasts a route update to neighbors which in turn they broadcast it to their neighboring nodes. All nodes receiving the update mark the base station as its parent and rebroadcast the update. The algorithm continues recursively with each node marking its parent as the first node from which it hears a routing update. All packets received or generated by a node are forwarded to its parent until they reach the base station [3].

As [7] and [3] show, the simplicity of this protocol makes it susceptible to all the attacks discussed in the previous section. Since routing updates are not authenticated, it is possible for any node to claim to be a base station and can become the parent of all nodes in the network. Authenticated routing updates will prevent an adversary from claiming to be a base station, but a powerful laptop class adversary can still carry out HELLO flood attacks by transmitting a high power message to all the nodes and by making every node to mark the adversary as the parent node.

An adversary interested in eavesdropping on, modifying, or suppressing packets in a particular area can do so by mounting a combined wormhole or sinkhole attack. The adversary first creates a wormhole between two colluding laptop-class nodes, one near the base station and one near the targeted area. The first node forwards authenticated routing updates to the second through the wormhole and rebroadcasts the routing update in the targeted area. Since the routing update through the wormhole will likely reach the targeted area considerably faster, the second node will create a large routing subtree in the targeted area with itself as the root [3]. As you can see in Figure 6 it might cause to selective forwarding attack.

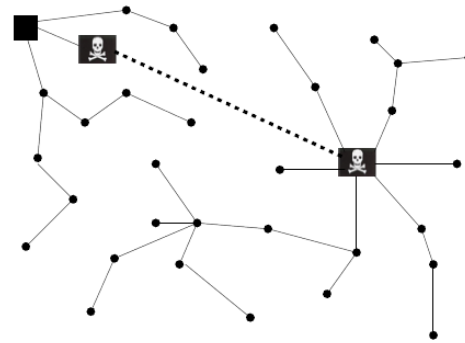


Fig. 6: A laptop-class adversary using a wormhole to create a sinkhole

### C. Geographic routing

Geographic and Energy Aware Routing (GEAR) [9] and Greedy Perimeter Stateless Routing (GPSR) [10] use nodes' positions and informed neighbor selection heuristics and also explicit geographic packet destinations to efficiently disseminate queries and route replies in the sensor network. GPSR uses greedy forwarding at each hop, routing each packet to the neighbor closest to the destination. During the routing, when some holes appear and greedy forwarding becomes impossible, GPSR recovers by routing around the perimeter of the void. One of the GPSR problems is that packets along a single flow will always use the same nodes for the routing of each packet, leading to uneven energy consumption.

GEAR attempts to solve this problem by weighting the choice of the next hop by both remaining energy and distance from the target. Every node has two different costs for accessing a destination. First, an estimated cost that is a mixture of residual energy in the power supply (battery) and its distance to the destination; and the second one, a learning cost that accounts for routing when the holes happens in the network. A hole is formed when there is no other node closer to the target other than itself. When a node receives a packet, it checks whether any of its neighbors is located closer to the target region. If there are more than one, the one that is closest to them target is chosen. If there is only one, it hoses that node. If there isn't any, then there is a hole and it picks one of its neighbors based on the learning cost function. So, both GEAR and GPSR protocols require location (and energy for GEAR) information to be exchanged between neighbors.

Location and cost information can be misrepresented. Attacks can be launched by an adversary node by just advertising to have maximum energy. For instance, in GEAR, an appropriate attack would be to always advertise maximum energy as well. An adversary can also

significantly increase her chances of success by mounting a Sybil attack. It can carry out Sybil attack by covering up the target node with multiple bogus nodes [3].

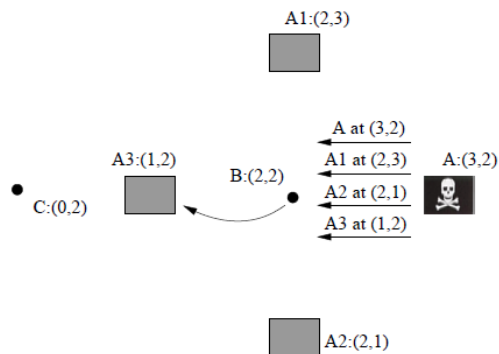


Fig. 7: The Sybil attack against geographic routing.

As shown in Figure 7, Adversary A at actual location (3,2) forges location advertisements for non-existent nodes A1, A2, and A3 as well as advertising her own location. After hearing these advertisements, if B wants to send a message to destination (0,2), it will attempt to do so through A3. This transmission can be overheard and handled by the adversary A. Once on that path, the adversary can mount a selective forwarding attack. In GPSR an adversary can forge location advertisements to create routing loops in data flows without having to actively participate in packet forwarding [3].

#### D. Rumor routing

Rumor routing [11] is a probabilistic protocol for matching queries with data events. Rumor routing offers a energy efficient alternative when the high cost of flooding cannot be justified. Rather than flooding the entire network to match information with interest (event), this protocol uses long lived packets called agents. When a source node observes an event it generates an agent. Agents pass through the whole network and propagate information about the local events to distant nodes. Agents carry information such as a list of events, next hop path to those events, hop count of those paths, a list of previously visited nodes and a Time To Live (TTL) field. On arriving at a new node the agent informs that node about the events it knows and adds to its event list. It decrements it's TTL field. If TTL is more than zero the node probabilistically selects the agent's next hop from its neighbors in the routing table minus the previously visited nodes by the agent.

In the same way the base station creates an agent to propagate the query into the network. A route from a base station to a source is established when a query agent arrives at a node previously traversed by an event agent that satisfies the query in the network [7].

As explain above, this protocol is dependent on nodes forwarding the agents properly. An adversary can mount a denial-of-service attack by removing event information carried by the agent or by refusing to forward agents entirely. Query or event information in agents can also be modified.

Laptop class attackers can carry out Sybil attacks and selective forwarding attacks [7].

Mote-class adversaries can mount a selective forwarding attack by extending tendrils in all directions to create a sinkhole. It creates tendrils by forwarding multiple copies of a received agent.

The easiest way to mount a selective forwarding attack is to be on the path of the data flow. Thus, the intersection of the query and events agents must occur downstream from the adversary. So, she will be "cut out" of the path of data flow [3].

#### V. SECURE SENSOR NETWORK ROUTING PROTOCOLS

Enforcing security in existing routing protocols through public key cryptographic mechanisms would either make them more complex or would consume the resources of tiny sensor devices.

According to these constraints, many secure routing protocols implement symmetric key cryptographic mechanisms to provide security. But this security is not complete because they consider only few of the design principles. For instance, SPINS and TinySec focus only on Prevention principle. They provide inadequate security in the presence of compromised nodes. As a preventive measure Secure Implicit Geographic Forwarding (SIGF) protocol chooses next hop dynamically and non-deterministically rather than maintaining routing tables [7].

On the other hands, Intrusion-Tolerant Routing protocol for Wireless Sensor Networks (INSENS) protocol uses multipath technique in order to make the network resilient to attacks.

Based on the [3], none of the proposed symmetric key based routing protocols incorporate all the three main design principles. These principles are Prevention, Detection or Recovery and Resilience. So to design and build a new protocol needs to consider all the discussed requirements.

Parno et al. has designed 'Secure Sensor Network Routing Protocol with a new asymmetric key based routing protocol and also security and efficiency as the central design parameters. The overhead and complexity of cryptographic mechanisms has been observed to be within acceptable limits [7].

DAWSEN (Defence mechanism Against Wormhole attacks in Wireless Sensor Networks) was introduced by Kaissi, R. and his group. They presented a defence mechanism against wormhole attacks in wireless sensor networks. Specifically, a simple routing tree protocol is proposed and shown to be effective in defending against wormhole attacks [2].

A proactive routing protocol based on the construction of a hierarchical tree where the base station is the root node, and the sensor nodes are the internal nodes of the tree. Each node receiving a request packet inserts a new entry in its “request list”. Then the node sends a reply packet and updates its “replay table”. Then the last 2 fields are set to zero. The node keeps listening to the transmitted reply packets, and increments the Num\_Rep field for each received packet. Now the source node sends for each entry in its reply list an equivalent accept packet. The node receiving an accept packet should check the source ID that should be the same as the NodeID in its replay table. If this is not the case, this will mean that this packet was stored by an attacker should be dropped. If not, the node updates its replay table by setting the “Recv\_accept” field to one and checks if the “Num\_reply” field in the accept packet is one value greater than “Num\_Rep” in the replay table of this node “  $\text{Num\_reply} = \text{Num\_Rep} + 1$  ”. If equation is verified, the node receiving the accept packet marks the originator of this packet as its parent, updates its routing table with the ID and the hop count of this parent and rebroadcasts a request packet with a hop count field incremented. And If the equation is not verified, a wormhole attack is detected by this node drop the received accept packet and add the ID of the originator of the accept packet to its NAP (Not Accepted Packets) table [2].

All of these routing protocols that were explained used a special simulation environment to run the simulations and evaluate the performance of the routing protocol. One of the most important of these simulators is network simulator 2 (ns2). It is a standard experiment environment in research community and creates some output files and collects statistical data synchronized from the sensor network [12].

## VI. CONCLUSIONS

Wireless Sensor Networks would be widely deployed in future mission-critical applications. As wireless sensor networks continue to grow and become more common, we expect that further expectations of security will be required of these wireless sensor network applications. One of these

considerations is security in routing protocol of wireless sensor network.

As I explained, some designs of sensor network routing protocols satisfy security goals of wireless sensor network. Link layer encryption and authentication mechanisms may be a reasonable first approximation for defense against mote-class outsiders, but cryptography is not enough to defend against laptop-class adversaries and insiders.

In contrast, according to my explanation, some currently proposed routing protocols for these networks are insecure. Table 1 shows briefly some attacks on these protocols.

Routing protocol	Selective Forwarding	Spoofed Attack	Sybil Attack	Sink Hole Attack	HELLO Attack
Directed diffusion	✓	✓	✓	✓	✓
TinyOS beaconing	✓	✓	✓	✓	✓
Geographic routing	✓	✓	✓	•	•
Rumor routing	✓	✓	✓	✓	•

Table 1: Summary of Attacks on routing protocols in Wireless Sensor Network

So, security problems at routing layer have to be resolved before their deployment in real world situations. A secure routing protocol should possess preventive measures against known attacks. Secure Sensor Network Routing protocol provides good security against all known attacks.

On detection of any suspicious activity of a malicious node recovery mechanisms should be triggered. Stability of the network should not be drastically disturbed even in the presence of the malicious node. Some secure routing protocols were explained and on implementing these protocols in particular operating system environment, it has been observed that the performance overhead is within acceptable limits compared to the level of security achieved.

## REFERENCES

- [1] J. Yick, B. Mukherjee, and D. Ghosal., *Wireless Sensor Network Survey*, 2008.
- [2] R. El-Kaissi, A. Kayssi, A. Chehab, and Z. Dawy., *DAWSEN: A Defence mechanism Against Wormhole attacks in Wireless Sensor Networks*.
- [3] C. Karlof and D. Wagner, *Secure Routing in Wireless Sensor Networks: Attacks and Countermeasures*, University of California at Berkeley.
- [4] J. Paul Walters, Z. Liang, W. Shi, and V. Chaudhary, *Wireless Sensor Network Security: A Survey*, Department of Computer Science Wayne State University.
- [5] S. Tripathy and S. Nandi, *Defense against outside attacks in wireless sensor networks*, Department of Information Technology, North Eastern Hill University, October 2007.

- [6] B. Sun, Y. Xiao, Ch. Chih Li, T. Andrew Yang, *Security co-existence of wireless sensor networks and RFID for pervasive computing*, Department of Computer Science, Lamar University, USA.
- [7] S. Shanmugham, *Secure Routing in Wireless Sensor Networks* Scholarly Paper Advisor: Dr. Jens-Peter Kaps.
- [8] C. Intanagonwivat, R. Govindan, and D. Estrin, *Directed diffusion: A scalable and robust communication paradigm for sensor networks*, in Proceedings of the Sixth Annual International Conference on Mobile Computing and Networks, August 2000.
- [9] Y. Yu, R. Govindan, and D. Estrin, *Geographical and energy aware routing: A recursive data dissemination protocol for wireless sensor networks*, University of California at Los Angeles Computer Science Department, May 2001.
- [10] B. Karp and H. T. Kung, *GPSR: greedy perimeter stateless routing for wireless networks*, in Mobile Computing and Networking, 2000.
- [11] D. Braginsky and D. Estrin, *Rumour routing algorithm for sensor networks*, in First ACM International Workshop on Wireless Sensor Networks and Applications, 2002.
- [12] J. Wang, *ns-2 Tutorial*, Multimedia Networking Group, The Department of Computer Science, 2004



# Skin-Color Based Videos Categorization

Rehanullah Khan<sup>1</sup>, Asad Maqsood<sup>1</sup>, Zeeshan Khan<sup>2</sup>, Muhammad Ishaq<sup>3</sup>, Arsalan Arif<sup>1</sup>

<sup>1</sup> Sarhad University of Science and Information Technology, Peshawar, Pakistan

<sup>2</sup> RWTH Aachen, The Chair of Human Language Technology and Pattern Recognition

<sup>3</sup> UET Mardan, Peshawar, Pakistan

## ABSTRACT

On dedicated websites, people can upload videos and share it with the rest of the world. Currently these videos are categorized manually by the help of the user community. In this paper, we propose a combination of color spaces with the Bayesian network approach for robust detection of skin color followed by an automated video categorization. Experimental results show that our method can achieve satisfactory performance for categorizing videos based on skin color.

Keywords: video categorization, skin detection in videos, color spaces

## 1. INTRODUCTION

Locating and tracking patches of skin-colored pixels through an image is a tool used in many face recognition and gesture tracking systems [13][8]. Skin information contributes much to object recognition [18]. One of the usage of skin color based tracking, locating and categorization could be blocking unwanted video contents on World Wide Web. On dedicated websites, people can upload videos and share it with the rest of the world. There are uploaded adult videos, which may not be allowed by the service providers. Therefore, how to effectively categorize and block such videos has been arousing a serious concern for the service providers.

The mostly used approach to contents blocking on the Internet is based on contextual keyword pattern matching technology that categorizes URLs by means of checking contexts of web pages or video names and then traps the websites [11][15]. This does not hold true for websites which allow uploading videos like Google Videos and YouTube, because the videos uploaded have different names from the contents they contain. Due to no automated process, the Google and YouTube rely on user's community. Therefore, an automated method to detect and categorize videos based on skin color will help the service providers and can provide control over the videos contents.

According to Smeulders et. al [14] color has been an active area of research in image retrieval, more than in any other branch of computer vision. The interest in color may be ascribed to the superior discriminating potentiality of a three

dimensional domain compared to the single dimensional domain of gray-level images [14].

The goal of our system is to categorize videos based on skin color. Depending on the percentage of skin in videos, the videos are flagged as Large-Skin-Videos (LSKIN), Partial-Skin-Videos (PSKIN) and No-Skin-Videos (NSKIN). The set of videos used in our experiments consists of 30 videos, collected and provided by video service provider. The service provider defines successful categorization as a true positive rate of above 70% because this would decrease the amount of manual work dramatically.

The remainder of the paper is organized as follows: Section 2 explains the previous work. Section 3 discusses color spaces used and the algorithm. Section 4 discusses experimental results and Section 5 concludes this paper.

## 2. PREVIOUS WORK

Singh et al. [13] discusses in detail different color spaces and skin detection. In their work, three color spaces; RGB, YCbCr and HSI are of main concern. They have compared the algorithms based on these color spaces and have combined them for face detection. The algorithm fails when sufficient non face skin is visible in the images. In [16], color spaces and their output results for skin detection are discussed. Furthermore, they state that excluding color luminance from the classification process cannot help achieving better discrimination of skin and non skin.

In [19] and [20], image filters based on skin information are described. The first step in their approach is skin detection. Maximum entropy modeling is used to model the distribution of skinness from the input image. A first order model is built that introduces constraints on color gradients of neighboring pixels. The output of skin detection is a gray scale skin map with the gray levels being proportional to the skin probabilities. There are false alarms when the background color matches human skin color. According to [3], a single color space may limit the performance of the skin color filter and that better performance can be achieved using two or more color spaces.

Jae et al. [10] discusses elliptical boundary model for skin color detection. To devise the appropriate model for

skin detection, they investigate the characteristics of skin and non-skin distributions in terms of chrominance histograms. They don't take the advantage of combining different color spaces. In [17], a method to detect body parts in images is presented. The algorithm is composed of content-based and image-based classification approaches. In the content-based approach, color filtering and texture analysis is used to detect the skin region in an image and its classification depends on the presence of large skin bulks. In the image-based approach, the color histogram and coherence vectors are extracted to represent the color and spatial information of the image.

According to [11] and [12], the selection of color space influences the quality of skin color modeling. The pixels belonging to skin region exhibit similar Cb and Cr chromatic characteristics, therefore, the skin color model based on Cb and Cr values can provide good coverage of all human races. Accordingly, despite their different appearances, these color types belong to the same small cluster in Cb-Cr plane. The apparent difference in skin colors perceived by viewers mainly comes from the darkness or fairness of the skin. These features are reflected on the difference in the brightness of the color, which is governed by Y component rather than Cb and Cr components. It provides an effective separation into luminance and chrominance channel and generates a compact skin chroma distribution. Yang et al. [18] have introduced a new Gamma Correction method to weaken the effects of illumination on images and a new RGB nonlinear transformation to describe the skin and non-skin distributions. Khan et al. [9][4][6] use face detection for adapting to the changing illumination circumstances for skin detection in videos. The authors in [5] introduce the usage of Decision Trees for pixel based skin detection and classification. Skin detection based on global seeds is introduced in [7].

### 3. SKIN COLOR MODELING

Color is a low level feature, which makes it computationally inexpensive and therefore suitable for real-time object characterization, detection and localization [11]. The main goal of skin color detection or classification for skin contents filtering is to build a decision rule that will discriminate between skin and non-skin pixels. Identifying skin colored pixels involves finding the range of values for which most skin pixels would fall in a given color space. This may be as simple as explicitly classifying a pixel as a skin pixel if Red, Green and Blue color channels have specific value range distribution. Other techniques use Neural Networks and Bayesian methods [11].

#### 3.1. RGB Color Space

In the RGB color space, each color appears in its primary spectral component of Red, Green and Blue. Images represented in the RGB space consist of three component images, one for each primary color. When fed into an RGB moni-

tor, these images combine on the phosphor screen to produce a composite color image. The RGB color space is one of the most widely used color spaces for storing and processing digital image [16]. However, the RGB color space alone is not reliable for identifying skin colored pixels since it represents not only color but also luminance [16]. Skin luminance may vary within and across persons due to ambient lighting, therefore, it is not suitable for segmenting skin and non-skin regions. Chromatic colors are more reliable and are obtained by eliminating luminance through nonlinear transformations [16].

#### 3.2. YCbCr

YCbCr is an encoded nonlinear RGB signal, commonly used by European Television Studios and for image compression work [16]. Color is represented by luma which is luminance and computed from nonlinear RGB constructed as a weighted sum of the RGB values and two color difference values Cb and Cr that are formed by subtracting luma from RGB Red and Blue components. The transformation simplicity and explicit separation of luminance and chrominance components make this color space attractive for skin color modeling [16].

#### 3.3. Skin Detection

For skin-color modeling, we construct a Bayesian network in the YCbCr and RGB color spaces. A Bayesian network is constructed from pixel triplet of the training skin colors.

A Bayesian network is also called a belief network and a directed acyclic graphical model. It is a representation for random variables and conditional independences within these random variables. The conditional independences are represented by Directed Acyclic Graph (DAG). More formally, a Bayesian network  $B = \langle N, A, \theta \rangle$  for skin color pixel (triplet) is a DAG  $\langle N, A \rangle$  with a conditional probability distribution for every node (collectively  $\theta$  for all nodes). A node  $n \in N$  in the graph  $G$  represents some random variable, and each edge or each arc  $A$  between nodes shows a probabilistic dependency. For learning Bayesian networks from specific datasets, data attributes are represented by nodes [1].

In a Bayesian network, the learner does not distinguish the skin and non-skin class variables from the attribute variables in data. As such, a network (or a set of networks) are created for skin color pixels that "best describes" the probability distribution of the training data. The problem of learning a Bayesian network can be stated as: Given a training set  $D = \{u_1, \dots, u_N\}$  of instances of  $U$ , find a network  $B$  that best matches  $D$ . Heuristic search techniques are used to find the best candidate in the space of possible networks. The search process relies on a scoring function that assesses the merits of each candidate network [2]. If we assume that for training, a Bayesian network  $B$  encodes a distribution  $P_B(A_1, \dots, A_n)$  from the training dataset with  $C$  classes, then

for testing, a classifier based on  $B$  returns the label  $c$  that maximizes the posterior probability  $P_B(c|a_1, \dots, a_n)$ . The network  $B$  can also be used to find out updated knowledge of the state of a subset of variables when other variables (the evidence variables) are observed.

### 3.4. Color-Space Intersection

The proposed skin categorization system starts with skin detection in videos based on the RGB color space. The detected skin pixels are passed to the YCbCr Bayesian detector. If the YCbCr skin detector confirms the pixels as skin pixels; the pixels are flagged as skin pixels. Depending on scenario and skin detected per frame, the video is flagged as LSKIN, PSKIN and NSKIN. Based on experiments, we set three rules for videos categorization: If the percentage of skin is greater than 15%, the video is flagged as LSKIN. If the skin percentage is greater than 3% and less than 15%, the video is flagged as PSKIN and NSKIN if less than 3%.

## 4. RESULTS

To evaluate the skin detection algorithm, we use a set of 30 challenging videos. Figure 1 shows example frames from these video sequences. The sequences span a wide range of environmental conditions. People of different ethnicity and various skin tones are represented. Sequences also contain scenes with multiple people and/or multiple visible body parts and scenes shot both indoors and outdoors, with moving camera. The lighting varies from natural light to directional stage lighting. Sequences contain shadows and minor occlusions. Videos in which background color matches the skin color are also present in the test set. Collected sequences vary in length from 100 frames to 1300 frames. These videos are divided into three categories depending on the amount of skin in video and serve as ground truth for the algorithm. Eleven videos are labeled as LSKIN, nine videos are labeled as PSKIN and ten are labeled as NSKIN.

On the testing set, the algorithm correctly identified 28 out of 30 videos as shown in table 1. Figure 2(a) shows an example skin detection on a single frame from Video 1. Figure 2(a) shows skin detection in the YCbCr color space. Figure 2(b) indicates the peaks related to the correct identification of skin in the entire Video 1. Figure 3 shows an example frame and skin detection in the RGB color space. This example frame is extracted from Video 2 which is correctly categorized as LSKIN based on skin presence.

The algorithm incorrectly reported two videos, Video 12 and Video 24 as LSKIN. The reason being the skin colored objects (false skin colors) present in these videos. Figure 4 is an example frame from Video 12 which is categorized as



Fig. 1. Examples frames from video sequences used for experimentation.



(a) Skin detection in the YCbCr color space (Video 1). Black shows skin.



(b) Skin graph showing peaks for Video 1 regarding correct identification of skin.

Fig. 2. Skin detection scenario.

Table 1. Result of videos categorization. Bold indicates wrong classification by the algorithm.

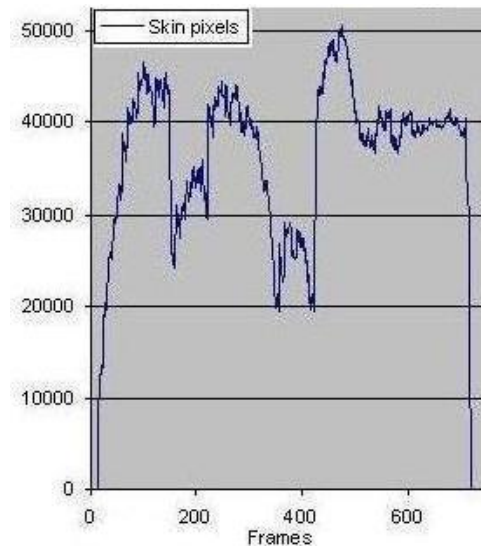
Video #	Ground Truth	Skin %	Result
1	LSKIN	22.10	LSKIN
2	LSKIN	19.50	LSKIN
3	LSKIN	17.60	LSKIN
4	LSKIN	28.66	LSKIN
5	LSKIN	25.23	LSKIN
6	LSKIN	17.67	LSKIN
7	LSKIN	24.57	LSKIN
8	LSKIN	29.90	LSKIN
9	LSKIN	16.30	LSKIN
10	LSKIN	25.20	LSKIN
11	LSKIN	29.90	LSKIN
<b>12</b>	<b>PSKIN</b>	<b>18.00</b>	<b>LSKIN</b>
13	PSKIN	07.99	PSKIN
14	PSKIN	07.55	PSKIN
15	PSKIN	07.70	PSKIN
16	PSKIN	11.97	PSKIN
17	PSKIN	12.77	PSKIN
18	PSKIN	10.81	PSKIN
19	PSKIN	09.99	PSKIN
20	PSKIN	08.99	PSKIN
21	NSKIN	02.25	NSKIN
22	NSKIN	02.88	NSKIN
23	NSKIN	01.99	NSKIN
<b>24</b>	<b>NSKIN</b>	<b>17.99</b>	<b>LSKIN</b>
25	NSKIN	02.77	NSKIN
26	NSKIN	01.99	NSKIN
27	NSKIN	02.78	NSKIN
28	NSKIN	01.00	NSKIN
29	NSKIN	01.22	NSKIN
30	NSKIN	02.50	NSKIN



Fig. 3. Skin detection in the RGB color space on a single frame from Video 2. Black shows skin.



(a) Skin detection example on a frame from Video 12. Black shows skin.



(b) Skin Peaks related to the incorrect skin detection in Video 12.

PSKIN but incorrectly reported as LSKIN. Figure 4(a) shows clothes and pig detected as skin. The female in Video 12 is wearing pink clothes that match the skin color. Figure 4(b) shows peaks related to the incorrect detection of clothes as skin for Video 12.

Figure 5 is an example frame from Video 24, which is also reported as LSKIN. Figure 5 shows that desert sand is detected as skin. When there is a sufficient match between the color of skin and the color of non-skin objects, the algorithm incorrectly reports it as skin. In such situation, color based skin categorization can be misleading. Texture analysis, use of semantics, and object recognition could help to distinguish skin colored background information from human skin color.

Fig. 4. Skin detection problems.

## 5. CONCLUSION

In this paper, we have developed an approach for categorization of videos based on skin color. We have tested our algorithm on 30 test sequences and achieved a true positive rate of over 90 %. In the next step, our goal is acquiring larger collections of videos in order to verify and improve the results.

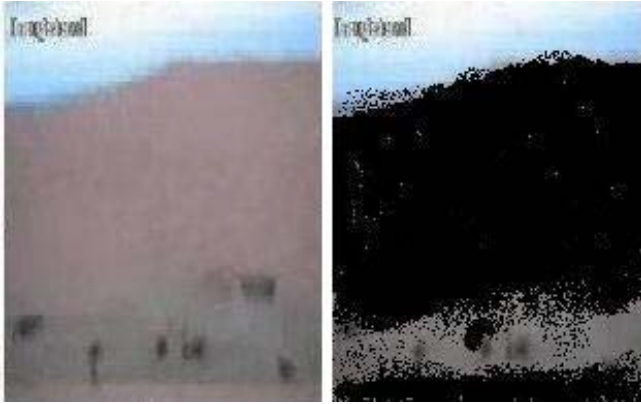


Fig. 5. Skin detection problems. Example frame from Video 24. Sand wrongly detected as skin. Black shows skin.

## 6. REFERENCES

- [1] Jie Cheng and Russell Greiner. Comparing bayesian network classifiers. In Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence, pages 101–110, San Francisco, CA, 1999. Morgan Kaufmann.
- [2] Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian network classifiers. *Mach. Learn.*, 29:131–163, November 1997.
- [3] Moheb R. Girgis, Tarek M. Mahmoud, and Tarek Abd-El-Hafeez. An approach to image extraction and accurate skin detection from web pages. In Proceedings of World Academy of Science, Engineering and Technology, pages 367–375, 2007.
- [4] Rehanullah Khan, A. Hanbury, and J. Stöttinger. Weighted skin color segmentation and detection using graph cuts. In Proceedings of the 15th Computer Vision Winter Workshop, pages 60–68, February 2010.
- [5] Rehanullah Khan, Allan Hanbury, and Julian Stöttinger. Skin detection: A random forest approach. In *ICIP*, pages 4613 – 4616, 2010.
- [6] Rehanullah Khan, Allan Hanbury, and Julian Stöttinger. Augmentation of skin segmentation. In *International Conference on Image Processing, Computer Vision, and Pattern Recognition*, pages 473–479, 2010.
- [7] Rehanullah Khan, Allan Hanbury, and Julian Stöttinger. Universal seed skin segmentation. In *International Symposium on Visual Computing*, pages 75–84, 2010.
- [8] Rehanullah Khan, Allan Hanbury, Julian Stöttinger, and Abdul Bais. Color based skin classification. *Pattern Recognition Letters*, 33(2):157 – 163, 2012.
- [9] Rehanullah Khan, Julian Stöttinger, and Martin Kampel. An adaptive multiple model approach for fast content-based skin detection in on-line videos. In *ACMMM, AREA workshop*, pages 89–96, 2008.
- [10] Jae Young Lee and Suk Yoo. An elliptical boundary model for skin color detection. In *ISST*, pages 579–584, 2002.
- [11] Jiann-Shu Lee, Yung-Ming Kuo, Pau-Choo Chung, and E-Liang Chen. Naked image detection based on adaptive and extensible skin color model. *PR*, 40(8):2261–2270, 2007.
- [12] Yung ming kuo, jiann-shu lee, and pau-choo chung. The naked image detection based on automatic white balance method. In *2006 ICS International Computer Conference*, pages 990–994, 2007.
- [13] Sanjay Kr. Singh, D. S. Chauhan, Mayank Vatsa, and Richa Singh. A robust skin color based face detection algorithm. *Tamkang Journal of Science and Engineering*, 6(4):227–234, 2003.
- [14] Arnold W.M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.
- [15] Julian Stöttinger, Allan Hanbury, Christian Liensberger, and Rehanullah Khan. Skin paths for contextual flagging adult videos. In *International Symposium on Visual Computing*, pages 303–314, 2009.
- [16] Vladimir Vezhnevets, Vassili Sazonov, and Alla Andreev. A survey on pixel-based skin color detection techniques. In *GraphiCon*, pages 85–92, 2003.
- [17] Shilin Wang, Hong Hui1, Sheng hong Li, Hao Zhang, Yong yu Shi, and Wen tao Qu. Exploring content-based and image-based features for nude image detection. In *Fuzzy Systems and Knowledge Discovery*, pages 324–328, 2005.
- [18] Jinfeng Yang, Zhouyu Fu, Tieniu Tan, and Weiming Hu. Skin color detection using multiple cues. In *Proceedings of the 17th International Conference on Pattern Recognition*, pages 632–635, 2004.
- [19] Huicheng Zheng, Mohamed Daoudi, and Bruno Jedynak. Adult image detection using statistical model and neural network. *Electronic Letters on Computer Vision and Image Analysis*, 4(2):1–14, 2003.
- [20] Huicheng Zheng, Mohamed Daoudi, and Bruno Jedynak. Blocking adult images based on statistical skin detection. *ELCVIA*, 4(2):1–14, 2004.

# Network Threat Ratings in Conventional DREAD Model Using Fuzzy Logic

Ak. Ashakumar Singh<sup>1</sup>, K.Surchandra Singh<sup>2</sup>

<sup>1</sup>Department of Computer Science, Thoubal College, Manipur University,  
Manipur-795138, India.

<sup>2</sup> Department of Computer Science, CMJ University, Shillong,  
Meghalaya, India.

## Abstract

One of the most popular techniques to deal with ever growing risks associated with security threats is DREAD model. It is used for rating risk of network threats identified in the abuser stories. In this model network threats needs to be defined by sharp cutoffs. However, such precise distribution is not suitable for risk categorization as risks are vague in nature and deals with high level of uncertainty. In view of these risk factors, the paper proposes a novel fuzzy approach using DREAD model for computing risk level that ensures better evaluation of imprecise concepts. Thus, it provides the capacity to include subjectivity and uncertainty during risk ranking. These threat parameters need to be frequently updated based on feedbacks from implementation of previous parameters. These feedbacks are always stated in the form of ordinal ratings, e.g. "high speed", "average performance", "good condition". Different people can describe different values to these ordinal ratings without a clear-cut reason or scientific basis. There is need for a way or means to transform vague ordinal ratings to more appreciable and precise numerical estimates. The paper transforms the ordinal performance ratings of some system performance parameters to numerical ratings using Fuzzy Logic.

**Keywords:** Fuzzy logic, threat rating, Transformation, DREAD Model.

## 1. Introduction

The computer network security remains a problem of great concern within information technology research area. Increase of network scale, development of advanced information technologies, and other factors enhance the number of possible targets for attacks against computer networks. These factors negatively influence upon the efficiency of the existing computer

networks security systems and enable research and development of *new protection models and technologies* [1].

[2]System security is one of the most significant issues in today's software society. Several security threats arise during software development process. Software failures, due to various vulnerabilities present in software, suggest essential presence of security in every phase of software development method. Microsoft's DREAD model is a popular approach for computing risk level of threats but it allows only crisp values [3]. Virtually every risk element can be characterized using two metrics, "Low, Medium, and High," or through "Ordinal Ranking". Therefore, most appropriate approach for defining risk level is using fuzzy logic. In this truth or validity of any statement becomes its degree of belongingness or membership.

This degree corresponds to a value to which an object is similar or compatible with the concept represented by fuzzy set. Truthfulness of a statement can be of various degrees which ranges from completely true, to partially true and then to completely false [4]. Moreover, fuzzy logic has linguistic values taken as words which can represent natural language for human reasoning during fuzzy rules construction. This means that these ratings have some elements of uncertainty, ambiguity or fuzziness.

When humans are the basis for an analysis, there must be a way to assign some rational value to intuitive assessments of individual elements of a fuzzy set. There is need to translate from human fuzziness to numbers that can be used by a computer.

Lofti A Zadeh introduced Fuzzy Set Theory (FST) in the early 1960's as a means of modeling the uncertainty, vagueness, and imprecision of human natural language. It was built on the basis that as the complexity of a system increases, it becomes more difficult and eventually impossible to make a precise statement about its behavior, eventually

arriving at a point of complexity where the fuzzy logic method born in humans is the only way to get at the problem.

[5] Described *Fuzzy Set Theory (FST)* as the extension of classical set theory. The basic idea is that the membership of a value to a set cannot only assume the two values “yes” or “no”, but can be expressed by gradual membership function within a range from zero to normally “1” in case of full membership degree. Membership function can assume several forms, and in practice triangular or trapezium forms are often used (Figure 1).

## 2. Problem Defined

The linguistics variables parameters of conventional DREAD Model viz. Damage potential (DP) of threat, Reproducibility (R) of the attack works, Exploitability (E) of threat, Affected User (A), and Discoverability(D) of the attacker are imprecise or fuzzy. The network threat parameters involved in the paper are respectively categorized as (1) Blind SQL Injection, (2) Login page SQL Injection, (3) Unencrypted login request, (4) Application error, (5) Inadequate account lockout, (5) Permanent cookie contains sensitive session information, (6) Session information not updated, (7) Unencrypted password parameter, and (8) Unencrypted view state parameter.

The ratings are in rough (imprecise, inexact or fuzzy) ranges, reflecting the variability in how each strategy could be implemented and the uncertainties involved in projecting the impacts of the strategies. For a meaningful numerical research, as stated in the introduction, these ordinal ratings need to be transformed to numerical ratings and this forms the thrust of the paper. That is, to transform opinion held by human beings, which would be "fuzzy" (e.g. low, mid-high performance) to being very precise (e.g. 15%, 80% performance), that is not "fuzzy" using fuzzy set theory [5], [6].

## 3. Theoretical Foundations

A fuzzy system is a system whose variable(s) range over states that are approximate. The fuzzy set is usually an interval of real number and the associated variables are linguistic variable such as “most likely”, “about”, etc. [5]. Appropriate quantization, whose coarseness reflects the limited measurement resolution, is inevitable whenever a variable represents a real-world attribute. Fuzzy logic consists of Fuzzy Operators such as “IF/THEN rules”, “AND, OR, and NOT” called the *Zadeh operators* [6].

The Membership Function is a graphical representation of the magnitude of participation of each input. It associates a weighting with each of the inputs that are processed, define

functional overlap between inputs, and ultimately determines an output response. Once the functions are inferred, scaled, and combined, they are defuzzified into a crisp output which drives the system. There are different memberships functions associated with each input and output response. Some features of different membership functions are: SHAPE - triangular is common, but bell, trapezoidal, haversine and, exponential have been used also; HEIGHT or magnitude (usually normalized to 1); WIDTH (of the base of function); SHOULDERING; CENTER points (centre of the member and OVERLAP (Figure 1) [9].

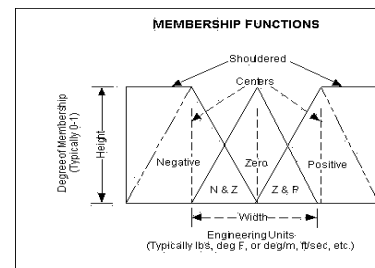


Fig. 1 Triangular membership function

The degree of fuzziness of a system analysis rule can vary between being very precise, that is not "fuzzy", to being based on an opinion held by a human, which would be "fuzzy." Being fuzzy or not fuzzy, therefore, has to do with the degree of precision of a system analysis rule.

The Degree of Membership (DOM) is the placement in the transition from 0 to 1 of conditions within a fuzzy set. The degree of membership is determined by plugging the selected input parameters into the horizontal axis and projecting vertically to the upper boundary of the Membership function(s) [9]. Fuzzy Variable includes words like red, blue, good and sweet are fuzzy and can have many shades and tints. A Fuzzy Algorithm is a procedure, usually a computer program, made up of statements relating linguistic variables. A Fuzzy Logic Control System-measures an input against a given situation and the system takes action automatically.

## 4. Methodology

The relative effectiveness of these threat ratings is summarizes as shown in Table 1 in terms of conventional five basic criteria: [2] (1)Damage Potential (DP) of threat, (2)Reproducibility (R) of the attack works, (3)Exploitability (E) of threat, (4)Affected User (A), and (5)Discoverability(D) of the attacker. In the table, the system performs between *medium to high* on practical reduction effectiveness, *high* in terms of economic efficiency, *medium to high* on economic equity for the poor and *medium to high* on immediate access flexibility.

## 5. Notations

- a Blind SQL Injection
- b Login page SQL Injection
- c Unencrypted login request
- d Application error
- e Inadequate account lockout
- f Permanent cookie contains sensitive session information
  
- g Session information not updated
- h Unencrypted password parameter
- i Unencrypted view state parameter
- me medium
- hi high
- lo low
- min Minimum
- Max Maximum
- Avg Average
- Temp Temperature
- Conc Concentration
- THR Threat

Table 1: Threat parameters ratings

Multi-objective Evaluation of the system					
Threat parameters	Ratings on Objectives (high = best)				
	DP	R	E	A	D
(a)	me-hi	hi	me-hi	me-hi	me
(b)	me-hi	me-hi	me-hi	me-hi	hi
(c)	lo-me	lo-me	me	hi	me-hi
(d)	lo-me	lo	me	hi	lo
(e)	lo-me	lo	me	me	lo-me
(f)	lo	me	hi	lo-me	lo
(g)	me	hi	lo-me	me-hi	me
(h)	me-hi	lo	me	lo	me-hi
(i)	hi	me	lo-me	me-hi	lo

## 6. Fuzzy Variables

In the paper, the adjectives describing the fuzzy variables and the range of threat are shown in Table 2. The Range of threat for the individual fuzzy variables is substituted in Table 1 to obtain Table 3.

Table 2: Fuzzy Variables and their ranges.

Fuzzy Variables	Range of threat %
High (hi)	75 – 100
Med-High (me-hi)	55 - 80
Med (Medium)(me)	35 - 60

Low-Med (lo-me)	15 - 40
Low (lo)	0 - 20

Table 3: Fuzzy Range of Performance for the individual fuzzy variables.

Multi-objective Evaluation of the system					
Threat parameters	Ratings on Objectives (high = best)				
	DP	R	E	A	D
(a)	55 - 80	75 - 100	55 - 80	55 - 80	35-60
(b)	55 - 80	55 - 80	55 - 80	55 - 80	75-100
(c)	15 - 40	15 - 40	35 - 60	75 - 100	55-80
(d)	15 - 40	0 - 20	35 - 60	75 - 100	0-20
(e)	15 - 40	0 - 20	35 - 60	35 - 60	15-40
(f)	0-20	35-60	75-100	15-40	0-20
(g)	35-60	75-100	15-40	55-80	35-60
(h)	55-80	0-20	35-60	0-20	55-80
(i)	75-100	35-60	15-40	55-80	0-20

## 7. Fuzzy Mapping

The fuzzy variables in Table 1, were transformed to numerical ratings using *Fuzzy Set Theory* as shown in Figures 2–6.

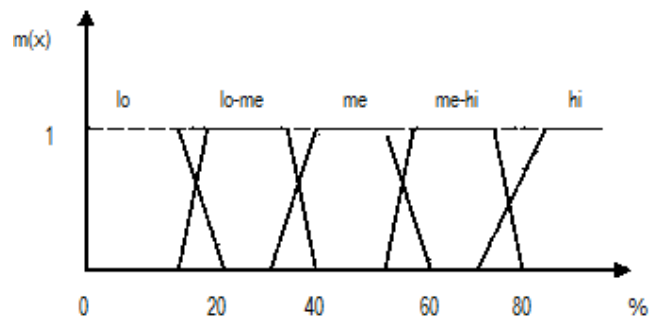


Fig. 2: Trapezoidal membership function

## 8. Aggregation of Fuzzy Scores

Using Figure 3, for each System parameters (SP)  $i$  and each criterion (CRIT)  $j$ ,

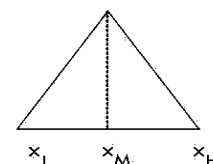


Fig. 3: Aggregation of Fuzzy Scores.

$i = 1, 2, 3, \dots, 9$  and  $j = 1, 2, 3, 4, 5$ .



For CRIT ( $j$ ) when  $SP(i, j) = x_L$  THEN  $SPTH R(i, j) = L$   
 For CRIT ( $j$ ) when  $SP(i, j) = x_M$  THEN  $SPTH R(i, j) = M$   
 For CRIT ( $j$ ) when  $SP(i, j) = x_H$  THEN  $SPTH R(i, j) = H$   
 Where, CRIT ( $j$ )  $\equiv$  Criterion  $j$  ( $j = 1, 2, 3, 4, 5$ )

$SP(i, j) \equiv$  System parameters  $i$  under Criterion  $j$   
 $SPTH R(i, j) \equiv$  System Threat parameters  $i$  under Criterion  $j$   
 Performance:

$$SPTH RSCORE(i) = \sum_j \frac{SP(i, j)}{5} \quad (1)$$

## 9. Membership Functions of the Fuzzy Sets

Using Aggregation methods for the fuzzy sets to reduce it to a triangular shape for the membership function, overlapping adjacent fuzzy sets were considered with the membership values shown in Figure 4.

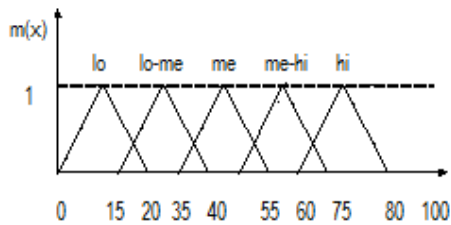
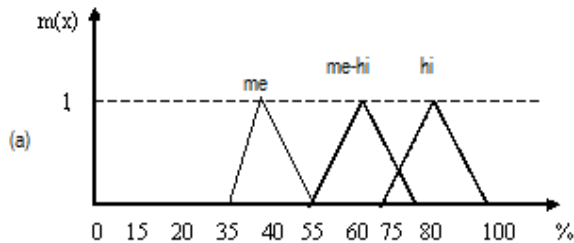
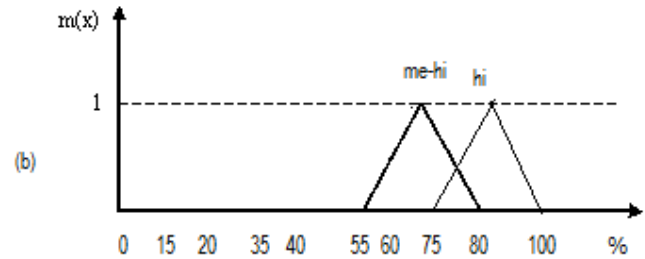


Fig. 4: Derived Triangular membership function

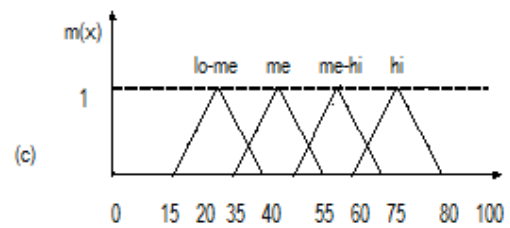
For the strategies and their performances, the membership functions shown in Figure 5 of the fuzzy sets were assigned.



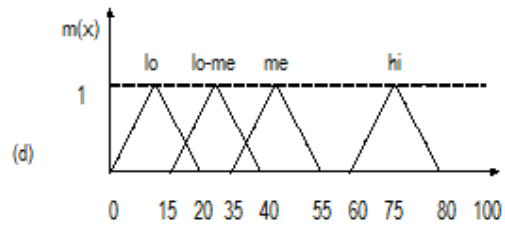
Criteria: (DP, E, A = med-hi; R = hi; D=me)



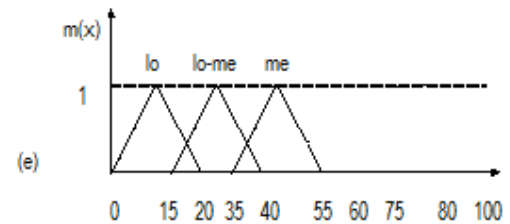
Criteria: (DP, R,E,A = me-hi; D=hi)



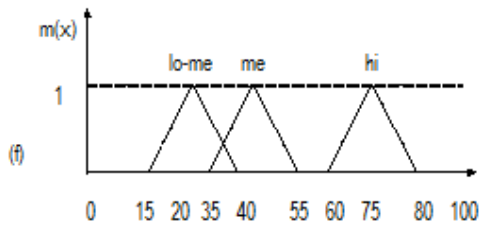
Criteria: (DP, R = lo-me; E = me; A = hi; D=me-hi)



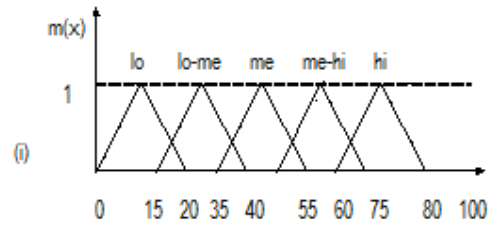
Criteria: (DP = lo-me; R, D = lo; E = me; A = hi)



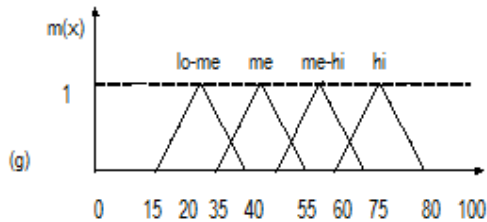
Criteria: (DP, D = lo-me; R = lo; E, A = me)



Criteria: (DP, D,A = lo-me; R = me; E=hi)



The ranges in figure 4 and figure 5 were aggregated to singletons. For the average performance of all the strategies, we have the fuzzy scaled rating as shown in figure 6.



Criteria: (DP, D=me; R =hi; E= lo-me; A = me-hi)

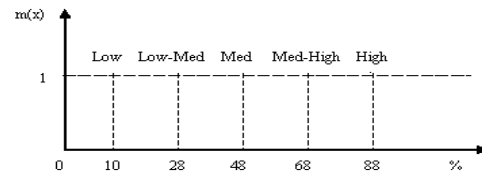
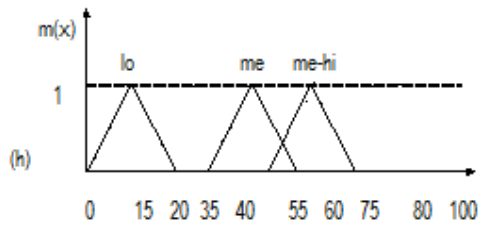


Fig. 6: Singleton aggregation of the ratings in table 1.



Criteria: (DP, D=me-hi; R,A=lo; E = me)

Criteria: (DP=hi; R=me; E=lo-me; A=me-hi; D =lo)

Fig. 5: Derived triangular membership functions for the System parameters.

From Figs. 2–6, the Membership Values assigned to each set of Universe of Discourse can be tabulated as shown in Table 4.

Table 4: Fuzzy threat ratings of Membership Values assigned to each set of Universe of Discourse.

Threat parameters	Criteria									
	DP		R		E		A		D	
(a)	Me-hi		hi		Me-hi		Me-hi		me	
	X	Y	X	Y	X	Y	X	Y	X	Y
	55	0	75	0	55	0	55	0	35	0
	68	1	88	1	68	1	68	1	48	1
	80	0	100	0	80	0	80	0	60	0
(b)	Me-hi		Med-hi		Me-hi		Me-hi		hi	
	X	Y	X	Y	X	Y	X	Y	X	Y
	55	0	55	0	55	0	55	0	75	0
	68	1	68	1	68	1	68	1	88	1
	80	0	80	0	80	0	80	0	100	0
(c)	Lo-me		Lo-me		me		hi		Me-hi	
	X	Y	X	Y	X	Y	X	Y	X	Y
	15	0	15	0	35	0	75	0	55	0

	28	1	28	1	48	1	88	1	68	1
	40	0	40	0	60	0	100	0	80	0
(d)	Lo-me		lo		me		hi		Lo	
	X	Y	X	Y	X	Y	X	Y	X	Y
	15	0	0	0	35	0	75	0	0	0
	28	1	10	1	48	1	88	1	10	1
	40	0	20	0	60	0	100	0	20	0
(e)	Lo-me		lo		me		me		Lo-me	
	X	Y	X	Y	X	Y	X	Y	X	Y
	15	0	0	0	35	0	35	0	15	0
	28	1	10	1	48	1	48	1	28	1
	40	0	20	0	60	0	60	0	40	0
(f)	Lo-me		lo		me		Lo-me		Lo	
	X	Y	X	Y	X	Y	X	Y	X	Y
	15	0	0	0	35	0	15	0	0	0
	28	1	10	1	48	1	28	1	10	1
	40	0	20	0	60	0	40	0	20	0
(g)	me		hi		Lo-me		Me-hi		me	
	X	Y	X	Y	X	Y	X	Y	X	Y
	35	0	75	0	15	0	55	0	35	0
	48	1	88	1	28	1	68	1	48	1
	60	0	100	0	40	0	80	0	60	0
(h)	Me-hi		lo		me		lo		Me-hi	
	X	Y	X	Y	X	Y	X	Y	X	Y
	55	0	0	0	35	0	0	0	55	0
	68	1	10	1	48	1	10	1	68	1
	80	0	20	0	60	0	20	0	80	0
(i)	hi		me		Lo-me		Me-hi		lo	
	X	Y	X	Y	X	Y	X	Y	X	Y
	75	0	35	0	15	0	55	0	0	0
	88	1	48	1	28	1	68	1	10	1
	100	0	60	0	40	0	80	0	20	0

(e)	15	0	35	35	15	20
(f)	15	0	35	15	0	13
(g)	35	75	15	55	35	43
(h)	55	0	35	0	55	29
(i)	75	35	15	55	0	36

## 10.Results

From the above figure 3, it is shown that  $x_L$ ,  $x_M$  and  $x_H$  are referred to as the Minimum performance, Average performance and Maximum Performance. Using equation (1), we can calculate the Average Scores of different LAN performance strategies for all the four criteria in respect of  $x_L$  referring to as the Minimum Performance (as shown in Table 5), in respect of  $x_M$  referring to as the Average Performance (as shown in Table 6), and in respect of  $x_H$  referring to as the Maximum performance (as shown in Table 7).

Table 5: Numerical transformation of Threat parameters for Minimum threat using fuzzy set theory.

Threat parameters	Multi-objective Evaluation of the System					
	Ratings on Objectives (high = best)					
	DP	R	E	A	D	Avg. Score
(a)	55	75	55	55	35	55
(b)	55	55	55	55	75	59
(c)	15	15	35	75	55	39
(d)	15	0	35	75	0	25

Table 6: Numerical transformation of Threat parameters for Medium Threat using fuzzy set theory.

Threat parameters	Multi-objective Evaluation of The System					
	Ratings on Objectives (high = best)					
	DP	R	E	A	D	Avg. Score
(a)	68	88	68	68	48	68
(b)	68	68	68	68	88	72
(c)	28	28	48	88	68	52
(d)	28	10	48	88	10	37
(e)	28	10	48	48	28	32
(f)	28	10	48	28	10	25
(g)	48	88	28	68	48	56
(h)	68	10	48	10	68	41
(i)	88	48	28	68	10	48

Table 7: Numerical transformation of the Threat parameters for Maximum Threat using fuzzy set theory.

Multi-objective Evaluation of System						
Threat parameters	Ratings on Objectives (high = best)					
	DP	R	E	A	D	Avg. Score
(a)	80	100	80	80	60	80
(b)	80	80	80	80	100	84
(c)	40	40	60	100	80	64
(d)	40	20	60	100	20	48
(e)	40	20	60	60	40	44
(f)	40	20	60	40	20	36
(g)	60	100	40	80	60	68
(h)	80	20	60	20	80	52
(i)	100	60	40	80	20	60

Table 8: Comparison between the ordinal fuzzy ratings and the transformed ratings on various criteria of LAN performance.

Ordinal (Fuzzy Ratings)	System parameters	Min Thr	Avg Thr	Max Thr
m-h	(a)	55	68	80
m-h	(b)	55	68	80
m-h	(c)	15	28	40

Similarly, for other fuzzy ratings of different System criteria, their comparisons can be found out.

Hence, their threat ratings can be shown such as  $x_L < x_M < x_H$ .

## 11. Conclusion

Fuzzy logic was used to transform ordinal System parameters of computer network threat ratings that are imprecise and fuzzy in nature to precise and defuzzified numerical ratings used in the analysis of threat ratings of different optimized threat parameters. The Technique used is the only way for solving any highly complex problem. The optimized system parameters will surely save the search time for technologies involved in the analysis of computer network threats.

### Reference:

- [1] Vladimir Gorodetski, Igor Kotenko and Oleg Karsaev, Multi-Agent technologies for computer network security: Attack simulation, intrusion detection and intrusion detection learning, *International Journal of Computer Systems Science and Engineering*, 18(4), 2003, 191-200.
- [2] Sonia, Archana Singhal and Banati, Fuzzy Logic Approach for threat prioritization in Agile Security Framework using DREAD Model, *International Journal of Computer Science Issues*, 8(4), July 2011, 182-190.
- [3] Howard M., and Leblanc D., Writing Secure Code, Second Edition, Microsoft Press, December 2002.
- [4] System Vulnerability Mitigation, The SANS Institute,

Dec-2003, [www.sans.org/system-vulnerability-mitigation\\_339-United States](http://www.sans.org/system-vulnerability-mitigation_339-United_States).

- [5] L.A. Zadeh, Fuzzy sets, *Information and Control*, 1965, 8, 338 – 353.
- [6] L.A. Zadeh, Toward a theory of fuzzy information granulation and its Centrality in human reasoning and Fuzzy logic, *International Journal of Soft Computing and Intelligence*, 90( 2), 1997, 111 – 127.
- [7] T.Sowell, *Fuzzy-Logic*, [http://www.fuzzy\\_logic.com/ch3\\_Htm](http://www.fuzzy_logic.com/ch3_Htm). (2005).
- [8] S.D. Kaehler, *Fuzzy Logic*, <http://www.seattlerobotics.org/encoder/mar98/fuz/flindex.html> (1998)
- [9] M.Kantrowitz, *Fuzzy-logic/part1*. [Online], 1997.

## Biography:



Ak. Ashakumar Singh graduated in Mathematics from Manipur University, Imphal and passed MCA in the year 2000 from the same varsity. He was awarded Ph.D. in the area of Computer Science from the Dept. of Mathematics of the same varsity in the year 2008. Then produced eight M.Phil scholars in Computer Science and now supervising three scholars leading to Ph.D. in Computer Science. The area of research is on Soft computing and related applications of computer science.



K. Surchandra Singh completed his Master of Computer Application (MCA) from Manipur University in the year 2000. Currently, he is a research scholar in the Department of Computer Science, CMJ University-India. He was working as faculty in the Institute of Cooperative Management, Imphal (India) since 2001 upto 2011, which has been conducting Post Graduate Diploma in Computer Application (PGDCA) affiliated to Manipur University, Canchipur (India). His research interest is in low cost network activities using Fuzzy logic principles.

# Using JQuery with Snort to Visualize Intrusion

Alaa El - Din Riad<sup>1</sup>, Ibrahim Elhenawy<sup>2</sup>, Ahmed Hassan<sup>3</sup> and Nancy Awadallah<sup>4</sup>

<sup>1</sup>Vice Dean for Students Affairs, Faculty of Computer Science and Information Systems, Mansoura University, Egypt  
Mansoura,,DK, 35513, Egypt

<sup>2</sup>Faculty of Computer Science and Information Systems, Zagazig University, Egypt  
Zagazig, Egypt

<sup>3</sup>Faculty of Engineering , Mansoura University , Egypt  
Mansoura,,DK, 35513, Egypt

<sup>4</sup>Faculty of Computer Science and Information Systems, Mansoura University, Egypt  
Mansoura,,DK, 35513, Egypt

## Abstract

The explosive growth of malicious activities on worldwide communication networks, such as the Internet, has highlighted the need for efficient intrusion detection systems. The efficiency of traditional intrusion detection systems is limited by their inability to effectively relay relevant information due to their lack of interactive / immersive technologies. Visualized information is a technique that can encode large amounts of complex interrelated data, being at the same time easily quantified, manipulated, and processed by a human user. Authors have found that the representations can be quite effective at conveying the needed information and resolving the relationships extremely rapidly. To facilitate the creation of novel visualizations this paper presents a new framework that is designed with using data visualization technique by using JQuery & Php for analysis and visualizes snort result data for user.

**Keywords:** *Intrusion Detection System, Snort rule Visualization techniques, , Php , JQuery.*

## 1. Introduction

Data visualization is a technique that has been used for a long time to represent information. Although old, yet powerful, its main outcome is that it allows the representation of data by different formats and shapes, each one highlighting a particular group of features.

Visualization represents a powerful link between the most dominant information-processing systems, the human brain and the modern computer. It is a key technology for extracting information, and therefore it is becoming more and more necessary in the field of Network Security. The power of network visualization goes beyond the simple "illustration" of network behavior to help the

analyst to discriminate between normal and abnormal activities. [1][2].

Using data visualization technique to support the result of snort (IDS) , we consider that PHP and JQuery as data visualization technique , we will deal with data of snort database to detect which data will be useful for network administrator to be visualized .

The framework introduced here is powerful because it is general, it can be applied to a wide domain of visualization problems. This research will assist users of visualization to explore, communicate, and understand their results.

The organization of this paper: next section discusses related research, section 3 discusses snort rules, and section 4 presents proposed system by using data visualization techniques for intrusion detection.

## 2. Related Research

**J.Blustein, C.Fu and D.L.Silver** presents proposed system that utilizes spatial hypertext workspace as the user interface could reduce the impact of high false alarm from IDS. This system may improvement the user's willingness to continuously monitor the system [3].

**R.F.Erbacher** discuss how user behavior can be exhibited within the visualization techniques, the capabilities provided by the environment, typical characteristics users should look out for (i.e., how

unusual behavior exhibits itself), and exploration paradigms effective for identifying the meaning behind the user's behavior [4].

**H.Koike and K.Ohno** propose a visualization system of a NIDS log named SnortView, which supports administrators in analyzing NIDS alerts much faster and much more easily. Instead of customizing the signature DB, they propose to utilize visualization to recognize not only each alert but also false detections [5].

**N.Rangaraju and M.Terk** describe a framework that is designed to simplify the process of building immersive visualization of structural analysis of building structures. They describe the components of the framework and describe two applications that were created to test their functionality [6].

**J.Peng, C.Feng and J.W.Rozenblit** proposed a hybrid intrusion detection and visualization system that leverages the advantages of current signature-based and anomaly detection methods. The hybrid intrusion detection system deploys these two methods in a two staged manner to identify both known and novel attacks.

When intrusion is detected, autonomous agents that reside on the system will automatically take actions against misuse and abuse of computer system, thus protecting the system from internal and external attacks [7].

**Y.Park and J.Park** presents Web Application Intrusion Detection System (WAIDS); an intrusion detection method based on an Anomaly Intrusion Detection model for detecting input validation attacks against web applications. Their approach is based on web application parameters which has identical structures and values. WAIDS derives a new intrusion detection method using generated profile from web request data in normal situation. By doing this, it is possible to reduce analysis time and false positives rate [8].

**R.U. Rehman** consider snort as an open source packet sniffer and logger that can be used as a lightweight Intrusion Detection System (IDS) to detect a variety of attacks and probes such as buffer overflows, stealth port scans, CGI attacks, and more. The Basic Analysis and Security Engine (BASE) displays and reports intrusions and attacks logged in the Snort database in a web browser for convenient analysis [9].

**A.Komlodi, J. R. Goodall and W.G. Lutters** report a framework for designing information visualization (IV) tools for monitoring and analysis activities. They studied ID analysts' daily activities in order to understand their routine work practices and the need for designing IV tools [10].

**K.Abdullah** presents new techniques to aid in network security using information visualization. Research contributions have been made in network data scaling and processing, port activity visualization, useful visualization showing a larger amount of information than textual methods, scaling port numbers and IP address for maximum use of screen space without occlusion, performing and using user study results to design an IDS alarm visualization tool [11].

**R.Erbacher, M.Garber** are attempting to improve the administrators ability to analyze the available data, make far more rapid assessments as to the nature of a given event or event stream, and identify anomalous activity not normally identified as such [12].

**K.Nyarko, T.Capers, etc.,** present a network intrusion visualization application with haptic integration, NIVA, which allows the analyst to interactively investigate as well as efficiently detect structured attacks across time and space using advanced interactive three-dimensional displays [13].

From previous studies we present our framework which be overcome on the problem of how to describe intrusion detection system results for network administrator.

### 3. Snort Rule

Snort uses a simple, lightweight rules description language that is flexible and quite powerful. Snort rules operate on network (IP) layer and transport (TCP/UDP) layer protocols [9].

#### 3.1 Rule Structure

Snort rules are divided into two logical sections as illustrated in Fig.1, the rule header and the rule options.



Fig. 1: Basic structure of Snort rules [9]

The rule header contains information about what action a rule takes. It also contains criteria for matching a rule against data packets. The options part usually contains an alert message and information about which part of the packet should be used to generate the alert message. The options part contains additional criteria for matching a rule

against data packets. A rule may detect one type or multiple types of intrusion activity. Intelligent rules should be able to apply to multiple intrusion signatures [9][14]. The general structure of a Snort rule header is shown in Fig. 2

Action	Protocol	Address	Port	Direction	Address	Port
--------	----------	---------	------	-----------	---------	------

Fig. 2: Structure of Snort rule header [9]

Fig. 3 describes snort rule sample:

```
alert tcp any any -> 192.168.1.0/24 111
(content:"|00 01 86 a5|"; msg: "mountd
access");
```

Fig. 3: Sample Snort Rule [14]

### 3.2 Rule classtype

Rules can be assigned classifications and priority numbers to group and distinguish them. Using the classification keyword is useful to prioritize intrusion detection data. The class type keyword, is found at the file (classification.config) which is included in the (snort.conf) file using the include keyword [9]. Each line in the (classification.config) file has the following syntax:

Config classification: name, description, priority.

The *name* is a name used for the classification. The *name* is used with the classtype keyword in Snort rules. The *description* is a short description of the class type. *Priority* is a number that shows the

default priority of the classification, which can be modified using a priority keyword inside the rule options. An example of this configuration parameter is as follows:

Config classification: DoS, Denial of Service Attack, 2

In the above line the classification is DoS (Denial of Service) and the priority is 2.

Fig. 4 illustrates rule uses default priority with the classification DoS:

```
alert udp any any -> 192.168.1.0/24 6838
(msg:"DoS"; \content: "server"; classtype:
DoS;)
```

Fig. 4: Using classtype in a rule [9]

In the next section we will use of the classification keyword in displaying Snort alerts by visualizing it through PHP & JQuery as visualization techniques.

## 4. Proposed System

This research aims to design a system for visualize intrusion detection by using PHP & JQuery as data visualization technique. The system introduces four components as showed in Fig. 5 and illustrated in detail in our previous paper [15].

This paper interest in "IDS Database", "Analysis Engine", "Visualized System" components which are described sequencing in section 4.1, 4.2 and 4.3.

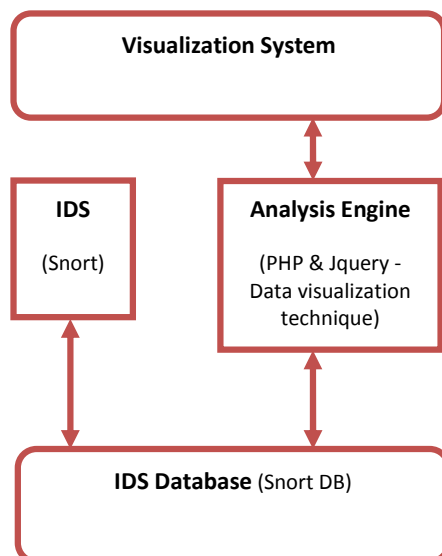


Fig. 5: Proposed System Structure

### 4.1 IDS (Snort) Database

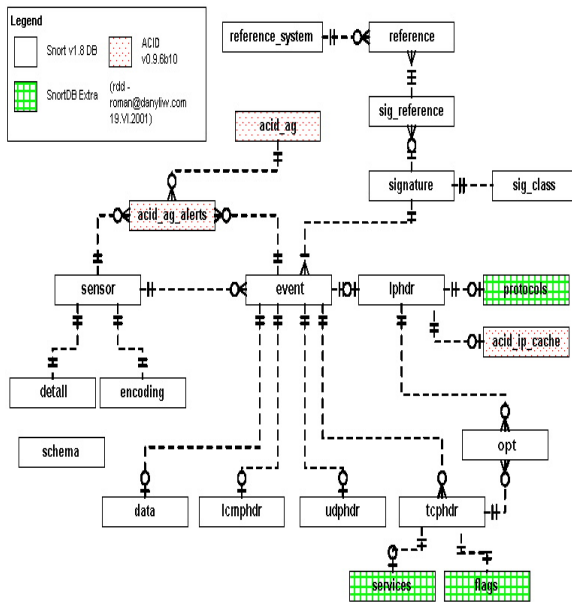


Fig. 6: Snort database schema [16]

The next table illustrates five tables of Snort database from which component related to (Snort or ACID) and its description.

Table 1: snort tables [16]

#### 4.2 Analysis Engine

This component responsible for retrieving data from snort database which be detected from snort (IDS) to be analyzed and processed it by PHP & JQuery.

The next figure (Fig.7) illustrated the relationship between previous tables in Table1.

Fig.8 is a screenshot of retrieving data from tables. There is a classification column which means signature classification: (Misc attack – Attempted user - Attempted recon- Attempted dos – Web application activity – Bad unknown)

After retrieving data from snort tables, we using JQuery & PHP to visualize signature classification as it will be showed in section 4.3.

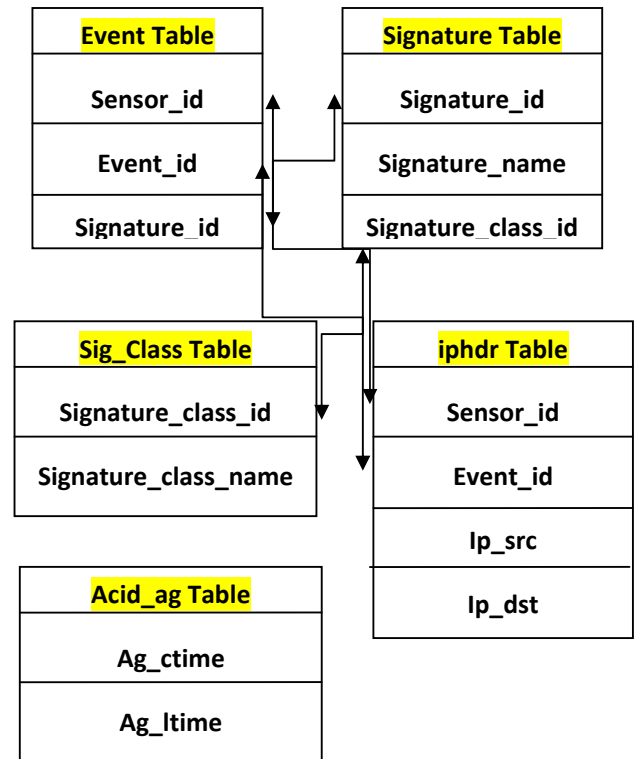


Fig.7: The relationship between used tables

Table	Component	Description
event	Snort	Meta-data about the detected alert
signature	Snort	Normalized listing of alert/signature names, priorities, and revision IDs
sig_class	Snort	Normalized listing of alert/signature classifications
iphdr	Snort	IP protocol fields
acid_ag	ACID	Meta-data for alert groups



Signature	Classification	Total #	Sensor #	Src.Addr	Dest.Addr	First	Last
[CVE][CVE] MISC UPNP malformed advertisement	Misc-attack	88855 (100%)	2	1	1	2002-11-06 04:32:28	2002-11-30 07:37:56
[url]BAD TRAFFIC loopback traffic	Bad-unknown	5(0%)	1	5	5	2002-11-06 06:01:15	2002-11-06 06:01:21
[CVE][CVE] SNMP broadcast trap	Attempted-recon	2(0%)	1	1	1	2002-11-09 11:13:21	2002-11-17 19:41:27
[bugtrap] [arachNIDS] WEB-CLIENT source via translate header	Web-application-activity	25(0%)	1	1	1	2002-11-12 17:42:03	2002-11-26 23:38:39
[bugtrap] [arachNIDS] EXPERIMENTAL WEB-CLIENT javascript host spoofing attempt	Attempted-user	1(0%)	1	1	1	2002-11-29 15:16:29	2002-11-29 15:16:29
SCAN Proxy (8080) attempt	Trojan-activity	2(0%)	1	1	1	2002-11-29 15:19:37	2002-11-29 15:19:37
WEB-IS scripts access	Attempted-dos	1(0%)	1	1	1	2002-11-29 15:27:02	2002-11-29 15:27:02

Fig.8: Use of the classification keyword in displaying Snort alerts [9]

The previous table included retrieved data from 5 tables which are illustrated and it's relation in fig.7: (Event, Signature, Sig\_Class, iphdr, Acid\_ag).

Researchers will use second column (**Classification**) from table in fig.8 to extract visualization system as it will be showed in fig.9 .

### 4.3 Visualization System

This component will be user interface for snort intrusion detection system result implemented by JQuery & PHP (Data Visualization Technique).

The next figure is an output from using PHP & JQuery as data visualization techniques after implementing data which was retrieved from (**Classification** column in fig.8) .This classification is according to signature (Misc-attack , Bad-unknown, Attempted-recon, Web-application-activity, Attempted-user , Trojan-activity, Attempted-dos).

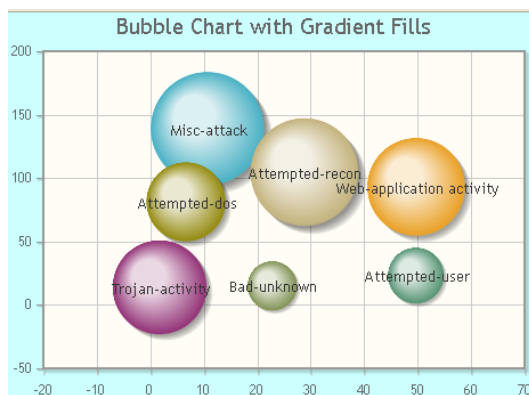


Fig. 9: Visualization System

### 4.4 Flowchart for proposed System

The next figure (fig.10) illustrates the flowchart of our proposed system which contains processes such as:

- Execute snort rules from (IDS component).
- Retrieving data from snort Database component.
- Using PHP & JQuery technique as analysis engine component.
- Extract new charts from analysis engine .

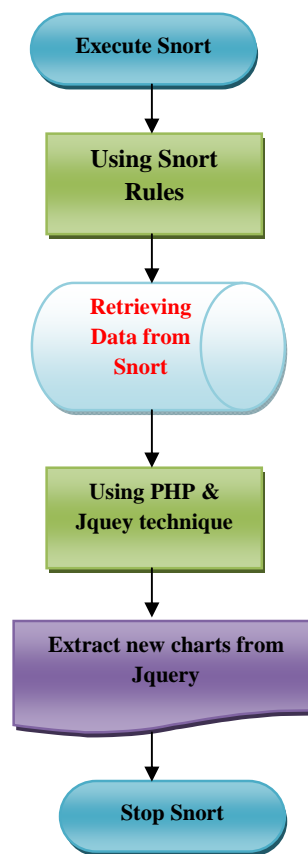


Fig.10: Flowchart for proposed System

## 5. Conclusion & Future Work

Intrusion detection is an information intensive and deeply analytic process that cannot be undertaken without the assistance of a computer.

Intrusion detection systems must handle masses of information (often in real-time) so as to report the abnormal use of networks and computer systems.

Our proposed system has proven to be effective for visually the intrusion which be detected by snort system.

In the future work we will compare between using visualization tools for intrusion detection such as NetGrok and using visualization techniques such as CSS & JQUERY.

## References

- [1] Nurbol, H. Xu, H.Yang, F.Meng, L. Hu, "A real-time intrusion detection security visualization framework based on planner-scheduler", IEEE ,2009 .
- [2] I. Onut , B.Zhu , A. Ghorbani, " A novel visualization technique for network anomaly detection" , proceedings of the 2nd Annual Conference on Privacy, Security and Trust (PST),p.167-174, 2004 .
- [3] J.Blustein, C.Fu, D.L.Silver, " Information Visualization for an Intrusion Detection System", ACM, 2005.
- [4] R.F.Erbacher, " Intrusion Behavior Detection Through Visualization", IEEE , 2003 .
- [5] H.Koike and K.Ohno , "SnortView: Visualization System of Snort Logs" IEEE , 2004 .
- [6] N.Rangaraju and M.Terk , " Framework for Immersive Visualization of Building Analysis Data " , IEEE , 2001 .
- [7] J.Peng, C.Feng and J.W.Rozenblit, "A Hybrid Intrusion Detection and Visualization System", IEEE , 2006 .
- [8] Y.Park and J.Park , " Web Application Intrusion Detection System for Input alidation Attack" , IEEE , 2008 .
- [9] R.U. Rehman " Intrusion Detection Systems with Snort Advanced IDS Techniques Using Snort, Apache, MySQL, PHP, and ACID" , Publishing as Prentice Hall PTR Upper Saddle River, New Jersey, 2003.
- [10] A.Komlodi, J.R. Goodall, W. G. Lutters, "An Information Visualization Framework for Intrusion Detection, 2004, IEEE
- [11] K.Abdullah , " Scaling and Visualizing Network Data to Facilitate in Intrusion Detection Tasks" ,Phd., School of Electrical and Computer Engineering ,Georgia Institute of Technology ,May 2006 .
- [12] R.Erbacher, M.Garber, "Visualization Techniques for Intrusion Behavior Identification " , 2004, IEEE.
- [13] K.Nyarko, T.Capers, C.Scott, K.Ladeji-Osias, " Network Intrusion Visualization with NIVA, an Intrusion Detection Visual Analyzer with Haptic Integration", 2002, IEEE.
- [14][http://petrinet.dvo.ru/pub/Vyatta/build-so/pkgs/vyatta-snort/debian/my/snort\\_rules.html](http://petrinet.dvo.ru/pub/Vyatta/build-so/pkgs/vyatta-snort/debian/my/snort_rules.html) , last visit on 29-11-2011
- [15] A.M.Riad, I. Elhenawy, A.Hassan, N.Awadallah, "Data Visualization Technique Framework for Intrusion detection", JCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, September 2011.
- [16][http://www.andrew.cmu.edu/user/rdanyliw/snort/acid\\_db\\_er\\_v102.html](http://www.andrew.cmu.edu/user/rdanyliw/snort/acid_db_er_v102.html) - last visit 03/10/2011

## Authors

**Alaa El - Din Riad**, Vice Dean for Students Affairs, Faculty of Computer and Information Sciences, Mansoura University

**Ibrahim Elhenawy**, Faculty of Computer and Information Sciences, Zagazig University

**Ahmed Hassan**, Department of Electrical Engineering, Faculty of Engineering, Mansoura University

**Nancy Awadallah** Researcher Assistant, Master in E-commerce Security 2008, Faculty of computer science & Information System - Mansoura University - Egypt

# Performance Evaluation of AODV and DSR with Varying Pause Time and Speed Time Over TCP and CBR Connections in VANET

Bijan Paul<sup>1</sup>, Md.Ibrahim<sup>2</sup> and Md.Abu Naser Bikas<sup>3</sup>

<sup>1</sup> Dept. of Computer Science & Engineering, Shahjalal University of Science & Technology  
Sylhet, Bangladesh

<sup>2</sup> Dept. of Computer Science & Engineering, Shahjalal University of Science & Technology  
Sylhet, Bangladesh

<sup>3</sup> Dept. of Computer Science & Engineering, Shahjalal University of Science & Technology  
Sylhet, Bangladesh

## Abstract

VANET (Vehicular Ad-hoc Network) is a new technology which has taken enormous attention in the recent years. Vehicular ad hoc network is formed by cars which are called nodes; allow them to communicate with one another without using any fixed road side unit. It has some unique characteristics which make it different from other ad hoc network as well as difficult to define any exact mobility model and routing protocols because of their high mobility and changing mobility pattern. Hence performance of routing protocols can vary with the various parameters such as speed, pause time, node density and traffic scenarios. In this research paper, the performance of two on-demand routing protocols AODV & DSR has been analyzed by means of packet delivery ratio, loss packet ratio & average end-to-end delay with varying pause time, speed time and node density under TCP & CBR connection.

**Keywords :** VANET; AODV; DSR; TCP; CBR; Packet Delivery Ratio; Average End-to-End Delay; Loss Packet Ratio

## 1. Introduction

VANET (Vehicular adhoc network) is an autonomous & self-organizing wireless communication network. In this network the cars are called nodes which involve themselves as servers and/or clients for exchanging & sharing information. This is a new technology thus government has taken huge attention on it. There are many research projects around the world which are related with VANET such as COMCAR [1], DRIVE [2], FleetNet [3] and NoW (Network on Wheels) [4], CarTALK 2000 [5], CarNet [6].

There are several VANET applications such as Vehicle collision warning, Security distance warning, Driver assistance, Cooperative driving, Cooperative cruise control, Dissemination of road information, Internet access, Map location, Automatic parking, and Driverless vehicles.

In this paper, we have evaluated performance of AODV and DSR based on TCP and CBR connection with varying pause time, speed time and also various network parameters and measured performance metrics such as packet delivery ratio,

loss packet ratio and average end-to-end delay of this two routing protocol and compared their performance. The remainder of the paper is organized as follows: Section 2 describes previous work related to performance evaluation of AODV and DSR and section 3 discusses about two unicast routing protocols AODV and DSR of VANET. Section 4 describes connection types like TCP and CBR. Section 5 presents performance metrics and the network parameters. Section 6 presents our implementation. Section 7 presents our decisions. We conclude in Section 8 and at the end add references.

## 2. RELATED WORK

There are several papers [7, 8, 9, 10, 11] related to performance evaluation of AODV and DSR .In [7], they measured packet delivery ratio, loss packet ratio and routing overhead using constant speed and pause time. Packet delivery ratio has been measured using variable speed and pause time in [8] .More broaden evaluation has been done in [9].The authors used TCP and UDP connection for performance comparison. Though there is a significant difference between TCP and CBR connection but comparison in between them is not yet analyzed. So we have focused on this two connection pattern based on different. We measured the performance of AODV and DSR with varying speed and constant pause time in [10].We used both high and low node density and observed the performance differences between TCP and CBR connection. Then we used varying pause time and constant speed in [11].We have observed that by changing the speed and pause time the performance varies between two connection. In this paper we have observed and analyzed the performance of AODV and DSR with varying pause time and speed.

## 3. Routing Protocols

An ad hoc routing protocol [12] is a convention, or standard, that controls how nodes decide which way to route packets in

between computing devices in a mobile adhoc network. There are two categories of routing protocol in VANET such as Topology based routing protocols & Position based routing protocols. Existing unicast routing protocols of VANET is not capable to meet every traffic scenarios. They have some pros and cons. We have already described it in our previous work [13]. We have selected two on demand routing protocols AODV & DSR for our simulation purpose.

### 3.1 AODV

Ad Hoc on Demand Distance Vector routing protocol [14] is a reactive routing protocol which establish a route when a node requires sending data packets. It has the ability of unicast & multicast routing. It uses a destination sequence number (DestSeqNum) which makes it different from other on demand routing protocols. It maintains routing tables, one entry per destination and an entry is discarded if it is not used recently. It establishes route by using RREQ and RREP cycle. If any link failure occurs, it sends report and another RREQ is made.

### 3.2 DSR

The Dynamic Source Routing (DSR) [15] protocol utilizes source routing & maintains active routes. It has two phases route discovery & route maintenance. It does not use periodic routing message. It will generate an error message if there is any link failure. All the intermediate nodes ID are stored in the packet header of DSR. If there has multiple paths to go to the destination DSR stores multiple path of its routing information.

AODV and DSR have some significant differences. In AODV when a node sends a packet to the destination then data packets only contains destination address. On the other hand in DSR when a node sends a packet to the destination the full routing information is carried by data packets which causes more routing overhead than AODV.

## 4.CONNECTION TYPES

There are several types of connection pattern in VANET. For our simulation purpose we have used CBR and TCP connection pattern.

### 4.1 Constant Bit Rate (CBR)

Constant bit rate means consistent bits rate in traffic are supplied to the network. In CBR, data packets are sent with fixed size and fixed interval between each data packets. Establishment phase of connection between nodes is not required here, even the receiving node don't send any acknowledgement messages. Connection is one way direction like source to destination.

### 4.2 Transmission Control Protocol (TCP)

TCP is a connection oriented and reliable transport protocol. To ensure reliable data transfer TCP uses acknowledgement, time outs and retransmission. Acknowledge means successful

transmission of packets from source to destination. If an acknowledgement is not received during a certain period of time which is called time out then TCP transmit the data again.

## 5.PERFORMANCE METRICS & NETWORK PARAMETERS

For network simulation, there are several performance metrics which is used to evaluate the performance. In simulation purpose we have used three performance metrics.

### 5.1 Packet Delivery Ratio

Packet delivery ratio is the ratio of number of packets received at the destination to the number of packets sent from the source. The performance is better when packet delivery ratio is high.

### 5.2 Average end-to-end delay

This is the average time delay for data packets from the source node to the destination node. To find out the end-to-end delay the difference of packet sent and received time was stored and then dividing the total time difference over the total number of packet received gave the average end-to-end delay for the received packets. The performance is better when packet end-to-end delay is low.

### 5.3 Loss Packet Ratio (LPR)

Loss Packet Ratio is the ratio of the number of packets that never reached the destination to the number of packets originated by the source.

## 6.OUR IMPLEMENTATION

For simulation purpose we used random waypoint mobility model. Network Simulator NS-2.34[16, 17] has been used. To measure the performance of AODV and DSR we used same scenario for both protocols.

### 6.1 Simulation Parameters

In our simulation, we used environment size 840 m x 840 m, node density 30 to 150 nodes with constant maximum speed 15 m/s and variable pause time 50 to 250 s. We did the Simulation for 200s with maximum 8 connections. The network parameters we have used for our simulation purpose shown in the table 1.

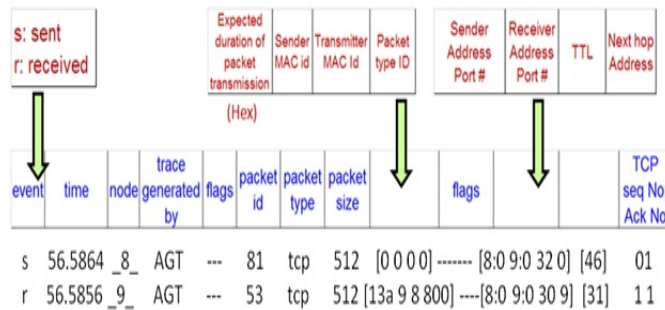
Table 1: Network Parameters

<i>Parameter</i>	<i>Value</i>
Protocols	AODV, DSR
Simulation Time	200 s
Number of Nodes	30, 90, 150
Simulation Area	840 m x 840 m
Speed Time	5, 10, 15, 20, 25 m / s
Pause Time	50, 100, 150, 200, 250 s
Traffic Type	CBR, TCP
Mobility Model	Random Waypoint
Network Simulator	NS 2.34

## 6.2 Performance Measurement Script

Generally, in NS-2 when we execute a program there creates two types of file trace file and nam file where nam file is used to visualize the simulation and trace file keep records of various interesting quantities such as each individual packets as its arrives, departs or is dropped at a link or queue by which we can measure a protocol performance.

### Trace file format



Awk script is required to analysis trace file for performance measure. To measure packet delivery ratio, loss packet ratio & average end-to-end delay of AODV and DSR we make two awk scripts. The scripts sudo codes are given below.

### PDR and LPR Measurement AWK Script

```
START
//initialization
SET nSentPackets to 0
SET nReceivedPackets to 0

IF $1 = "s" AND $4 = "AGT" THEN
    INCREMENT nSentPackets
ENDIF

IF $1 = "r" AND $4 = "AGT" THEN
    INCREMENT nReceivedPackets
ENDIF

COMPUTE rPacketDeliveryRatio as nReceivedPackets /
nSentPackets * 100

COMPUTE lpr as ( nSentPackets-nReceivedPackets ) /
nSentPackets ) * 100

PRINT nSentPackets
PRINT nReceivedPackets
PRINT rPacketDeliveryRatio
PRINT lpr
END
```

### END-TO-END DELAY AWK Script

```
START
//initialization
SET seqno to -1
SET count to 0

IF $4 = "AGT" AND $1 = "s" AND seqno < $6 THEN
    COMPUTE seqno as $6
ENDIF

IF $4 = "AGT" AND $1 == "s" THEN
    COMPUTE start_time[$6] as $2
ELSE IF $7 = "tcp" AND $1 = "r" THEN
    COMPUTE end_time[$6] as $2
ELSE IF $1 = "D" AND $7 = "tcp" THEN
    COMPUTE end_time[$6] as -1
ENDIF

FOR X = 1 to seqno
    IF end_time[X] > 0 THEN
        COMPUTE delay[X] as end_time[X] - start_time[X]
        INCREMENT count
    ELSE
        COMPUTE delay[i] as -1
    ENDIF
END FOR

FOR X = 1 to seqno
    IF delay[X] > 0 THEN
        COMPUTE n_to_n_delay as n_to_n_delay + delay[X]
    ENDIF
END FOR

COMPUTE n_to_n_delay as n_to_n_delay / count *
1000

PRINT n_to_n_delay

END
```

## 6.3 Simulation Results Analysis

The performance of AODV & DSR has been analyzed with varying pause time 50s to 250s and speed time 5 to 25 m/s for number of nodes 30, 90, 150 under TCP & CBR connection. We measure the packet delivery ratio, loss packet ratio & average end-to-end delay of AODV and DSR and the simulated output has shown by using graphs.

## 6.4 Graphs

Based on the simulation result we have generated the graph which shows the differences between AODV and DSR. The graphs are given below.

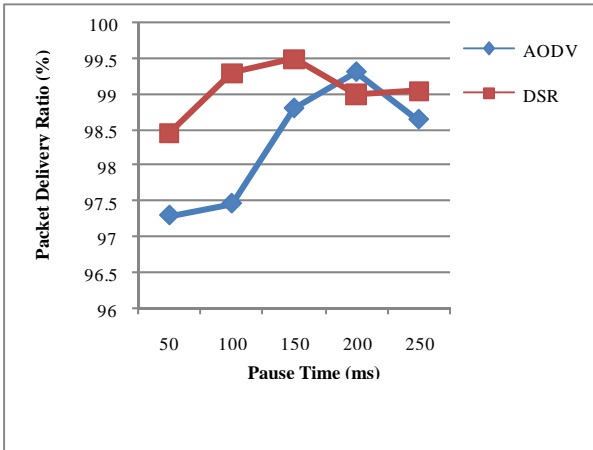


Figure1. PDR (w.r.t. Pause) of 30 nodes using TCP

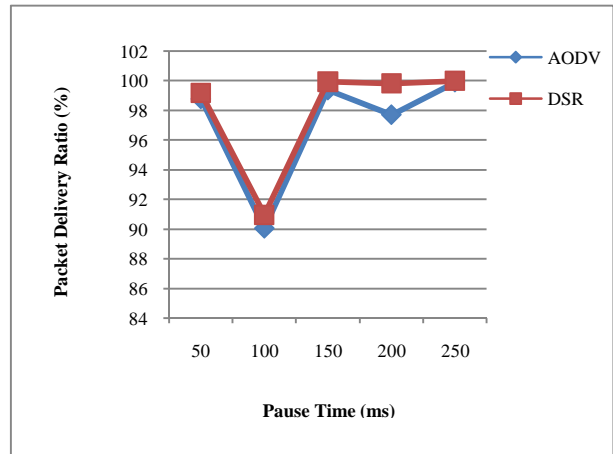


Figure 4. PDR (w.r.t. Pause) of 30 nodes using CBR

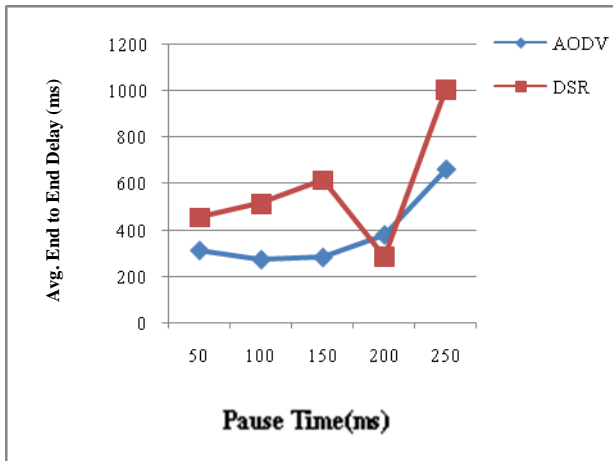


Figure 2. Avg.E2E delay (w.r.t. Pause) of 30 nodes using TCP

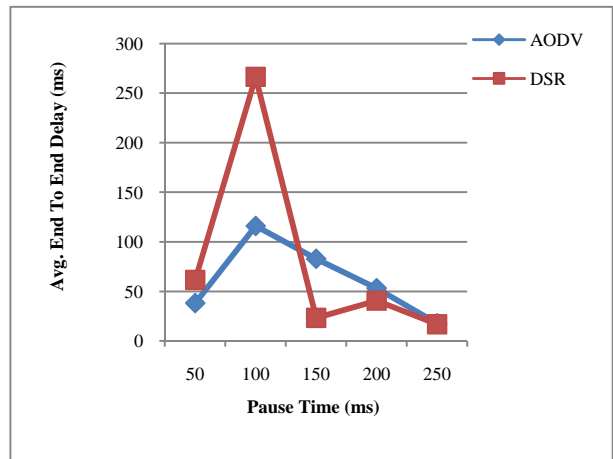


Figure 5. Avg.E2E delay (w.r.t. Pause) of 30 nodes using CBR

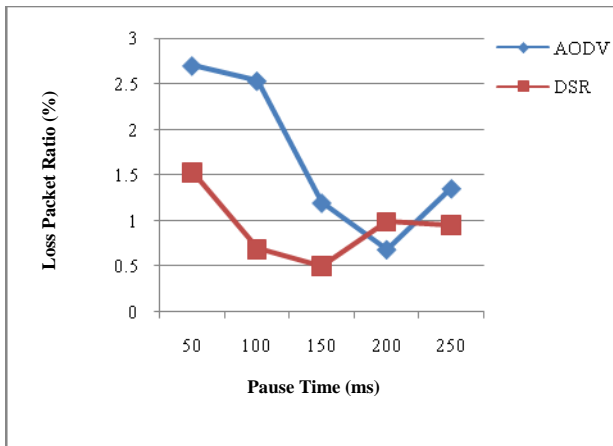


Figure 3. LPR (w.r.t. Pause) of 30 nodes using TCP

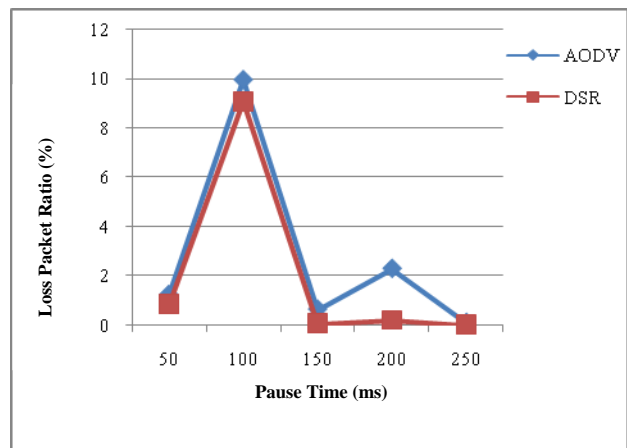


Figure 6. LPR (w.r.t. Pause) of 30 nodes using CBR

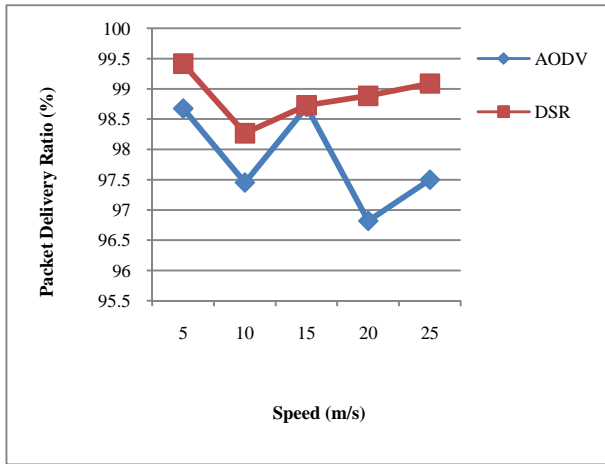


Figure 7. PDR (w.r.t. Speed) of 30 nodes using TCP

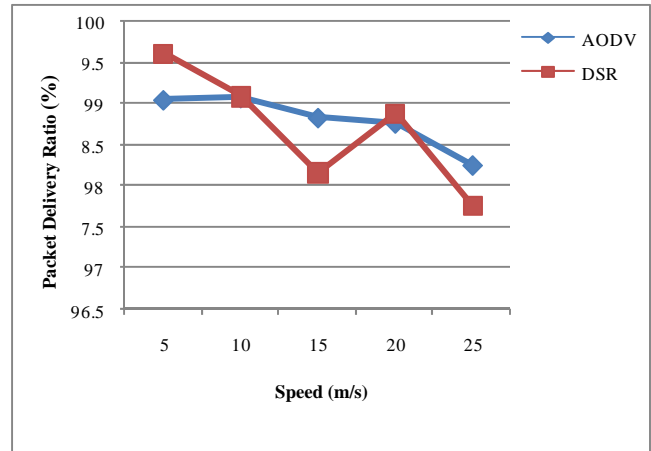


Figure 10. PDR (w.r.t. Speed) of 30 nodes using CBR

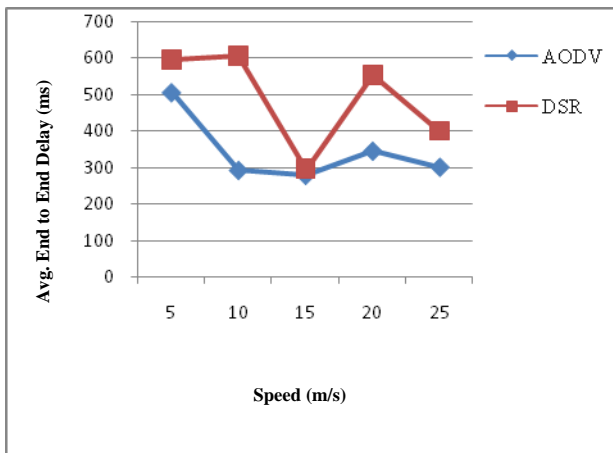


Figure 8. Avg. E2E delay (w.r.t. Speed) of 30 nodes using TCP

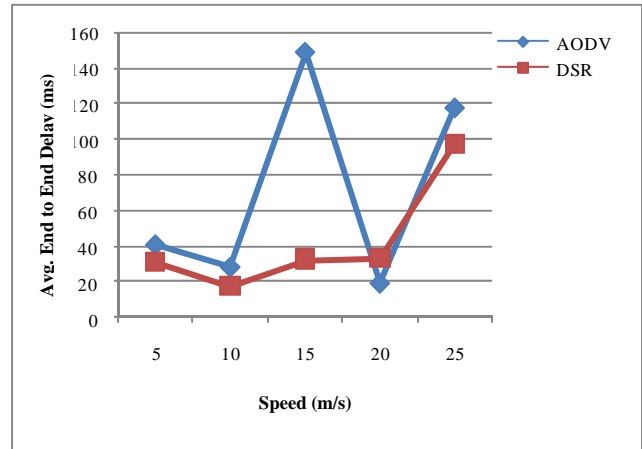


Figure 11. Avg. E2E delay (w.r.t. Speed) of 30 nodes using CBR

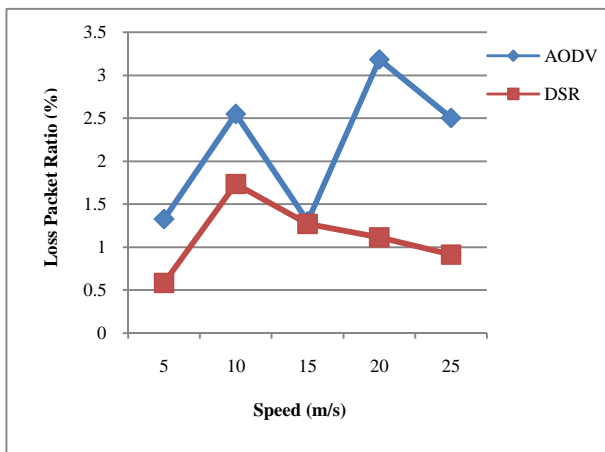


Figure 9. LPR (w.r.t. Speed) of 30 nodes using TCP

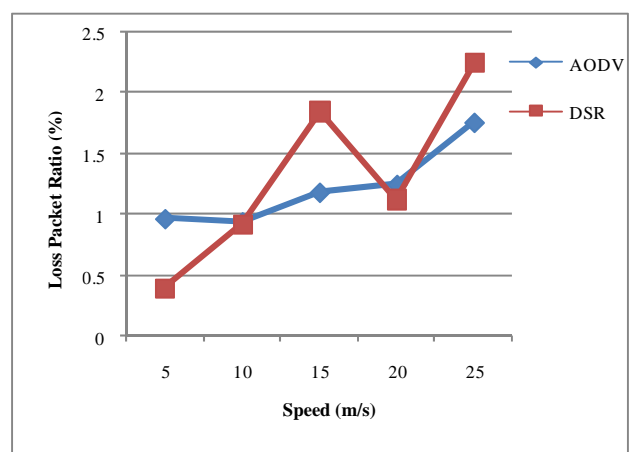


Figure 12. LPR (w.r.t. Speed) of 30 nodes using CBR

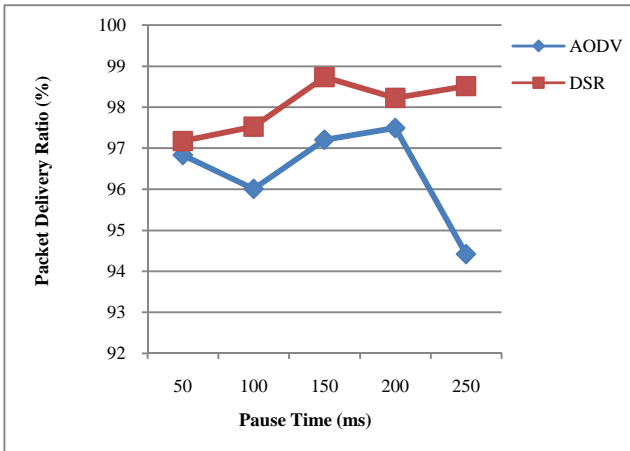


Figure 13. PDR (w.r.t. Pause) of 90 nodes using TCP

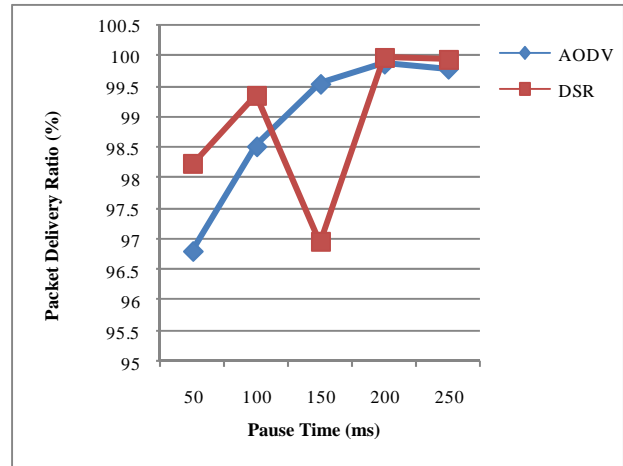


Figure 16. PDR (w.r.t. Pause) of 90 nodes using CBR

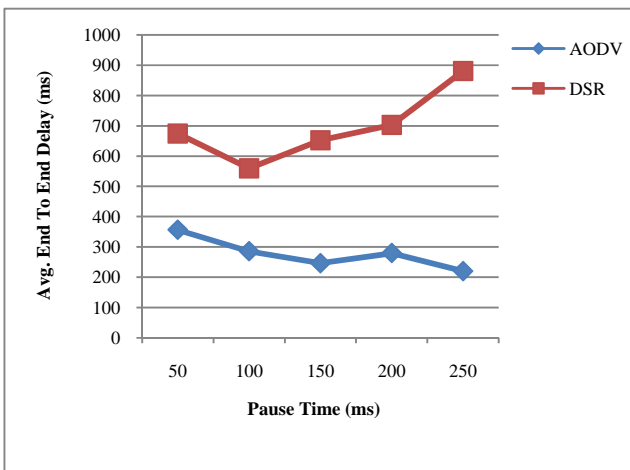


Figure 14. Avg.E-2-E delay (w.r.t. Pause) of 90 nodes using TCP

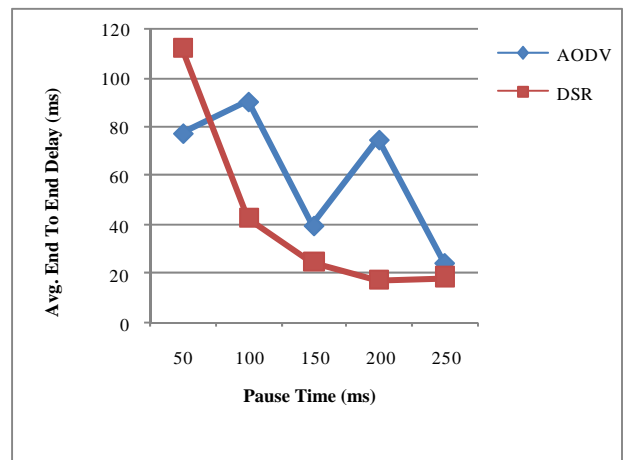


Figure 17. Avg.E-2-E delay (w.r.t. Pause) of 90 nodes using CBR

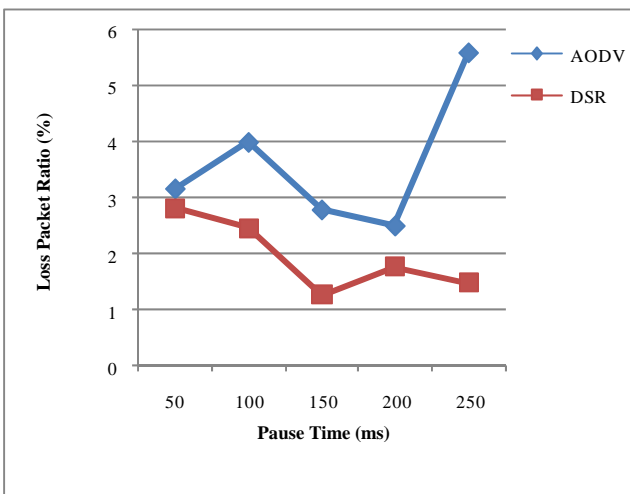


Figure 15. LPR (w.r.t. Pause) of 90 nodes using TCP

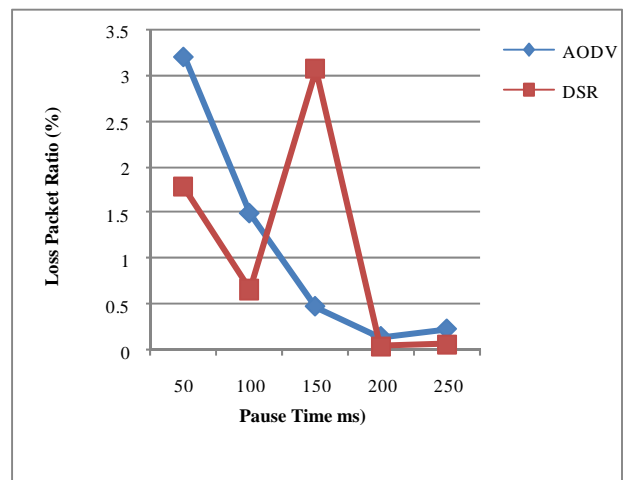


Figure 18. LPR (w.r.t. Pause) of 90 nodes using CBR



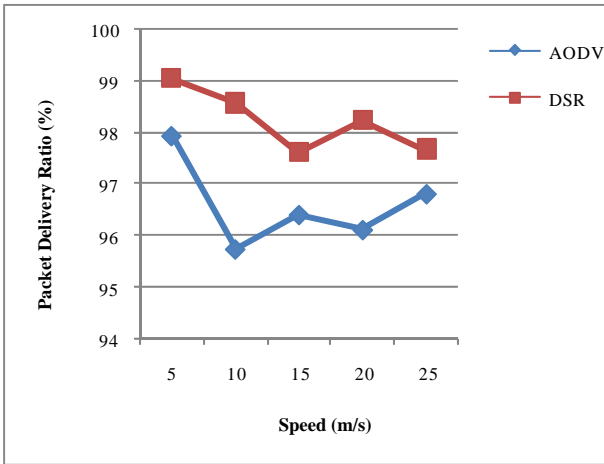


Figure 19. PDR (w.r.t. Speed) of 90 nodes using TCP

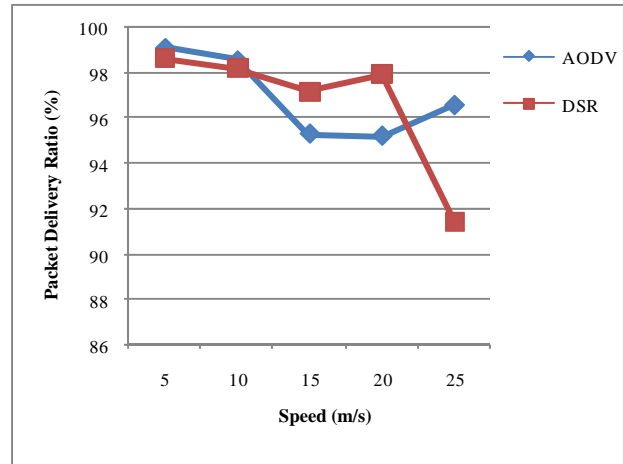


Figure 22. PDR (w.r.t. Speed) of 90 nodes using CBR

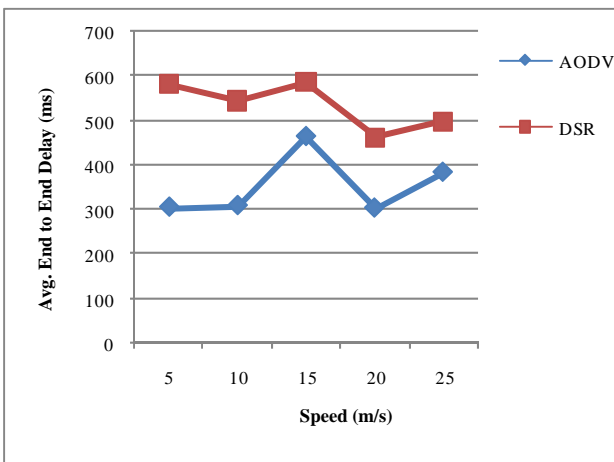


Figure 20. Avg.E2E delay (w.r.t. Speed) of 90 nodes using TCP

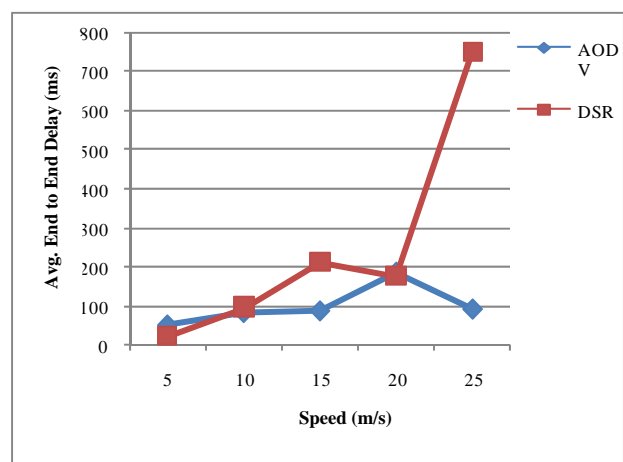


Figure 23. Avg.E2E delay (w.r.t. Speed) of 90 nodes using CBR

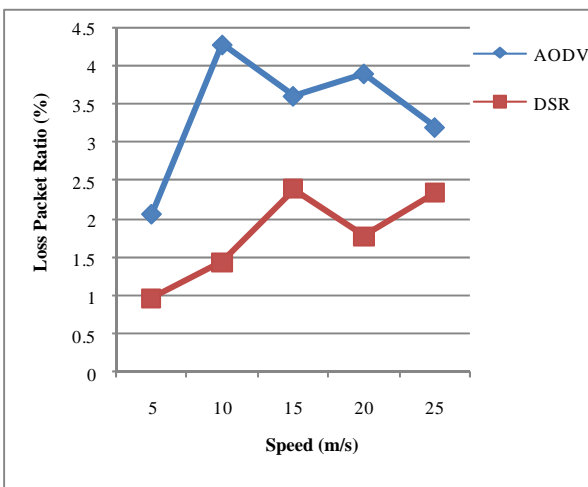


Figure 21. LPR (w.r.t. Speed) of 90 nodes using TCP

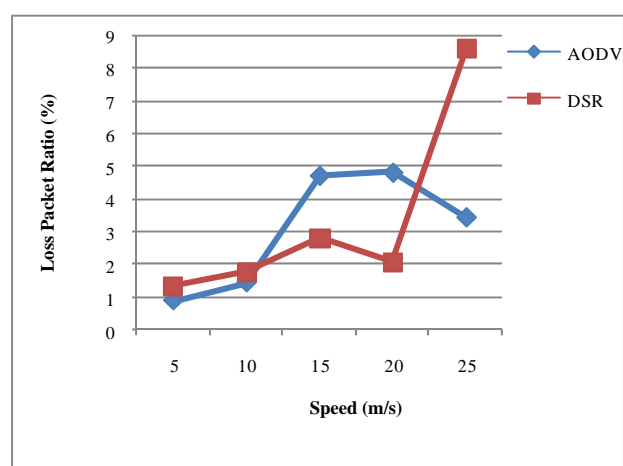


Figure 24. LPR (w.r.t. Speed) of 90 nodes using CBR

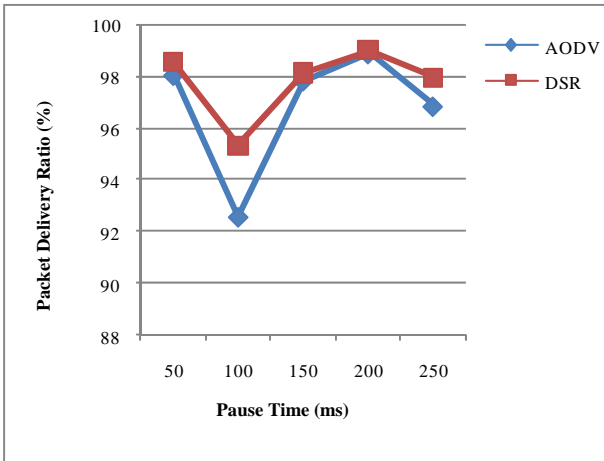


Figure 25. PDR (w.r.t. Pause) of 150 nodes using TCP

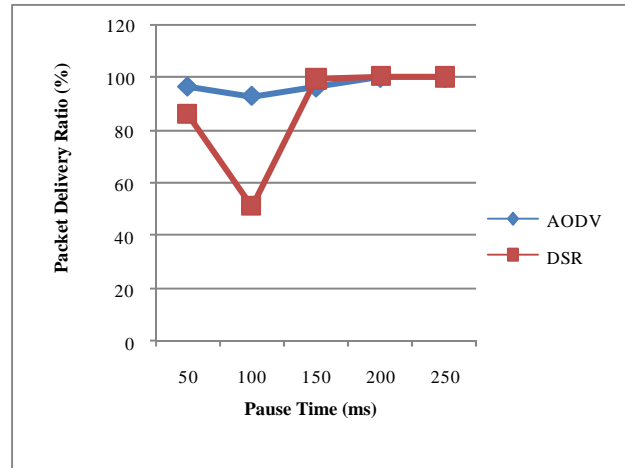


Figure 28. PDR (w.r.t. Pause) of 150 nodes using CBR

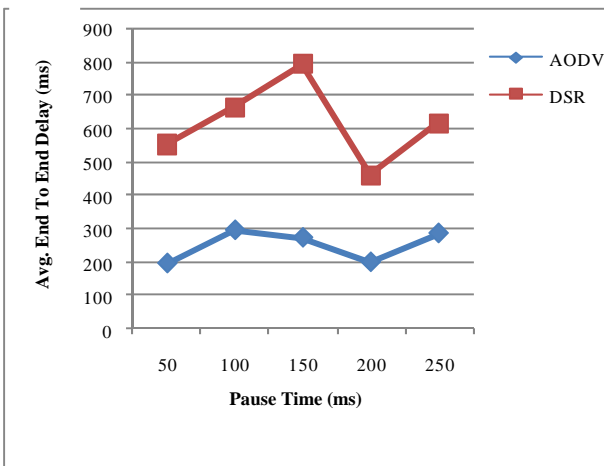


Figure 26. Avg.E2E delay (w.r.t. Pause) of 150 nodes using TCP

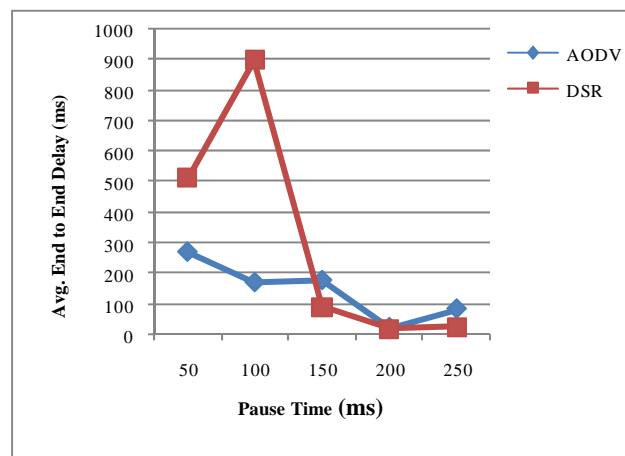


Figure 29. Avg.E2E delay (w.r.t. Pause) of 150 nodes using CBR

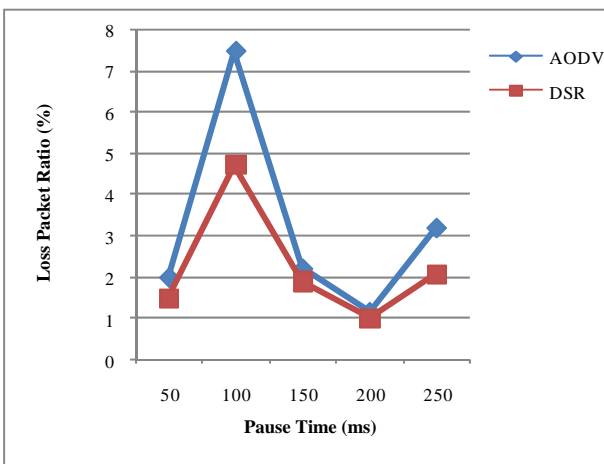


Figure 27. LPR (w.r.t. Pause) of 150 nodes using TCP

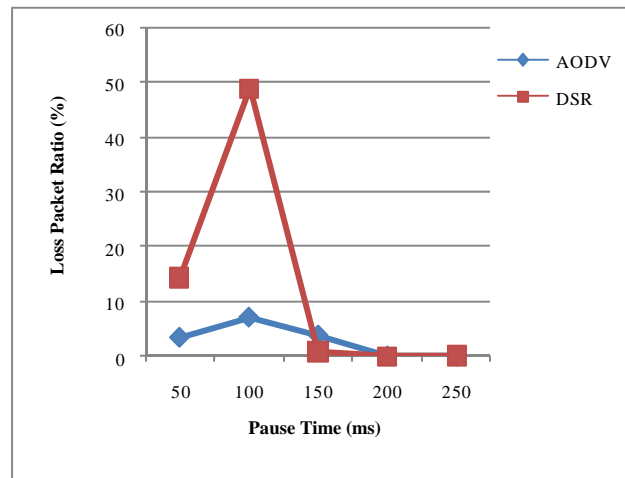


Figure 30. LPR (w.r.t. Pause) of 150 nodes using CBR

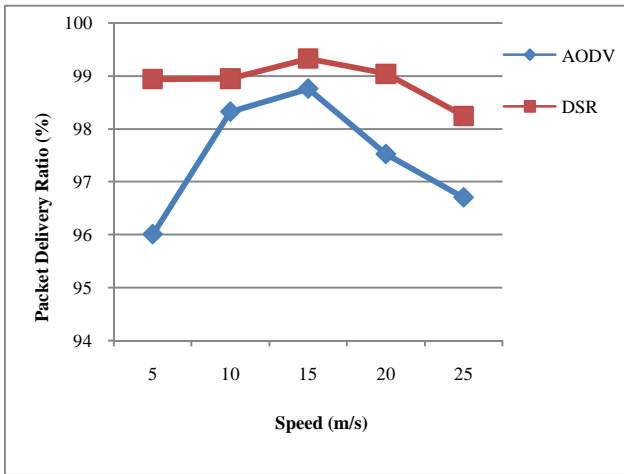


Figure 31: PDR (w.r.t. Speed) of 150 nodes using TCP

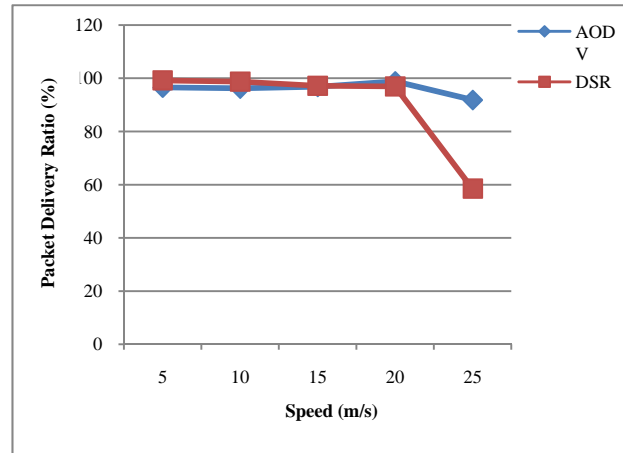


Figure 34: PDR (w.r.t. Speed) of 150 nodes using CBR

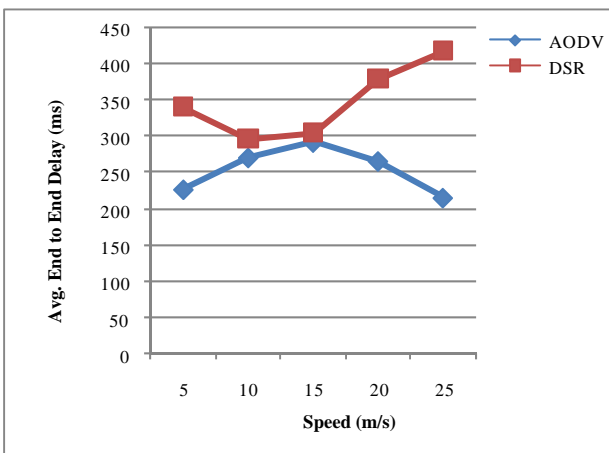


Figure 32: Avg.E2E delay (w.r.t. Speed) of 150 nodes using TCP

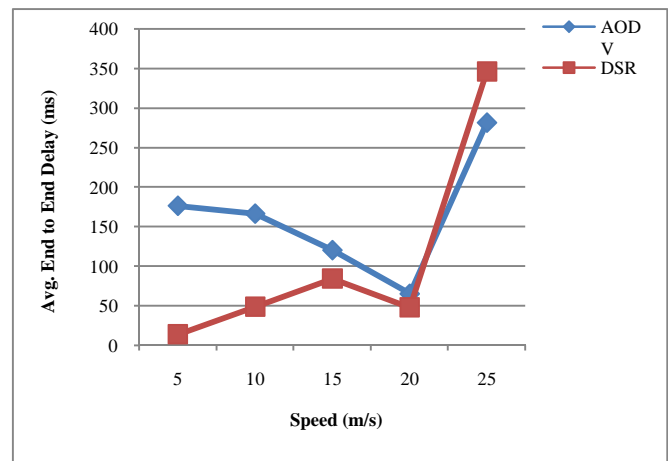


Figure 35: Avg.E2E delay (w.r.t. Speed) of 150 nodes using CBR

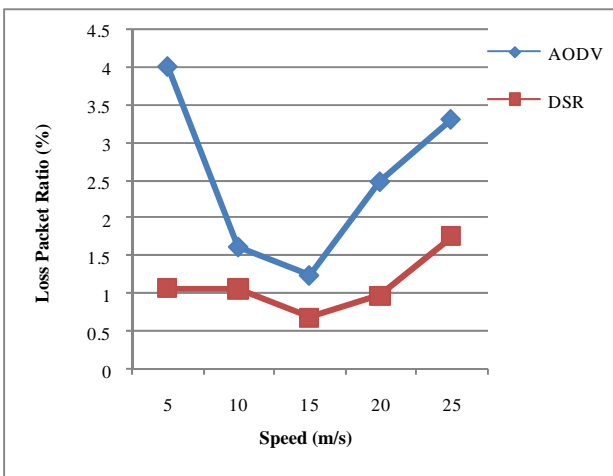


Figure 33: LPR (w.r.t. Speed) of 150 nodes using TCP

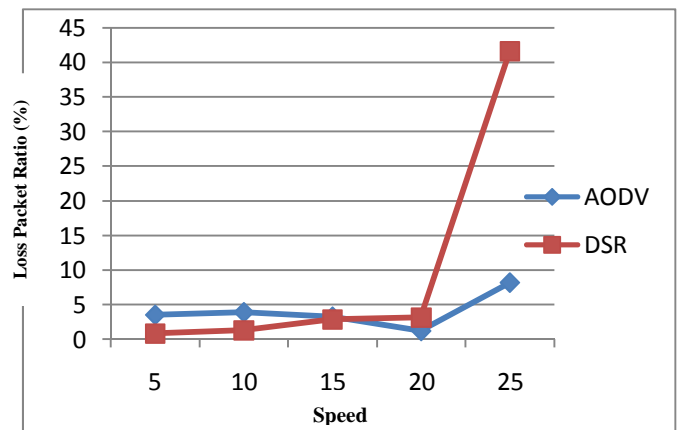


Figure 36: LPR (w.r.t. Speed) of 150 nodes using CBR

### 6.5 Analysis Table

After analysis of AODV and DSR we define a standard for simulation results. We consider 30 nodes as low density, 90 nodes as average density and 150 nodes as high density. We also consider 5 m/s as low speed, 15 m/s as average speed and 25 m/s as high speed.

The standard for PDR values (approx.) defines below

**For speed & pause time:**

High:  $\geq 98\%$

Average: 96% to 97%

Low:  $\leq 95\%$

The standard for E-to-E values (approx.) defines below

**For pause time:**

High:  $\geq 351\text{ms}$

Average: 151ms to 350ms

Low:  $\leq 150\text{ms}$

**For speed time:**

High:  $\geq 150\%$

Average: 51% to 150%

Low:  $\leq 50\%$

The standard for LPR values (approx.) define below

**For pause time:**

High:  $> 2\%$

Average: 1% to 2%

Low:  $< 1\%$

**For speed time:**

High:  $> 3\%$

Average: 1.5% to 3%

Low:  $< 1.5\%$

Our simulation area considered is  $840 \times 840$  and simulation run time is 200 seconds. Speed has been varied from 5m/s to 25 m/s. Pause time has been varied from 50s to 250s. Based on our standard we can summarize the following differences between AODV and DSR based on our estimated parameters.

**Pattern analysis of 30 nodes using TCP connection**

From our experimental analysis we observe that for TCP connection using pause time as a parameter in low mobility low pause time the packet delivery ratio (PDR) is average for AODV and high for DSR. In that scenario average end to end delay (E-To-E) is average for AODV and high for DSR. The loss packet ratio for TCP connection is high for AODV and average for DSR. If the pause time is high the PDR for both routing protocols is high. E-To-E for both protocols is high. LPR of DSR is low but for AODV it is average. On the other hand, using speed as a parameter in low mobility low speed the packet delivery ratio for both protocols is high. In that scenario average end to end delay (E-To-E) is high, the loss packet ratio is low for both routing protocol. But in low mobility high speed, the PDR for AODV is average but high for DSR. E-To-E for both protocols is high. LPR of AODV is average. But for DSR it is low.

**Pattern analysis of 30 nodes using CBR connection**

We observe that for CBR connection using pause time as a parameter in low mobility low pause time the packet delivery

ratio (PDR) of CBR for both routing protocols is high. In that scenario average end to end delay (E-To-E) is low for both protocols. The loss packet ratio is average for AODV and low for DSR. If the pause time is high the PDR for both routing protocols is high. E-To-E is low for both routing protocols. LPR of DSR is low. But for AODV it is low. On the other hand, using speed as a parameter in low mobility low speed the packet delivery ratio for both protocols is high. In that scenario average end to end delay (E-To-E) and the loss packet ratio is low for both routing protocol. But in low mobility high speed, the PDR for AODV is high but average for DSR. E-To-E for both protocols is low. LPR is average for both routing protocols.

**Pattern analysis of 150 nodes using TCP connection**

Pause time as a parameter in high mobility low pause time PDR for both protocols is high. In that scenario average end to end delay (E-To-E) is average for AODV and high for DSR. The LPR is average for both protocols. If the pause time is high the PDR for both routing protocols is average. E-To-E is average for AODV and high for DSR. LPR is high for AODV and DSR.

On the other hand, using speed as a parameter in high mobility low speed, PDR of AODV is average but high for DSR. Though, E-To-E for AODV & DSR is high. LPR is low for DSR and high for AODV. If the speed is high AODV performs average and DSR performs high. E-To-E is high for both routing protocol. LPR of AODV is high but for DSR it is average.

**Pattern analysis of 150 nodes using CBR connection**

We observe that for CBR connection using pause time as a parameter in high mobility low pause time the packet delivery ratio (PDR) of CBR it is average for AODV and low for DSR. E-To-E for AODV is average but it is high for DSR. The loss packet ratio is high for both protocols. If the pause time is high the PDR for AODV and DSR using CBR is high. E-To-E and LPR is low for both routing protocols.

On the other hand, using speed as a parameter in high mobility low speed the packet delivery ratio for AODV is average but high for DSR, Though E-To-E and LPR for AODV is high but low for DSR. If the speed is high the PDR for AODV and DSR is low. E-To-E is high for both routing protocol. LPR of AODV and DSR is high for CBR connection.

## 7. OUR DECISIONS

After performance analysis of AODV & DSR by using decision table we declare our decision.

TABLE-2: PDR, E-2-E AND LPR WITH RESPECT TO LOW MOBILITY & LOW PAUSE TIME FOR TCP & CBR CONNECTIONS

Protocols	Packet Delivery Ratio		Avg. End to End Delay		Loss Packet Ratio	
	TCP	CBR	TCP	CBR	TCP	CBR
AODV	Avg	High	Avg	Low	High	Avg
DSR	High	High	High	Low	Avg	Low

TABLE-3: PDR, E-2-E AND LPR WITH RESPECT TO LOW MOBILITY & HIGH PAUSE TIME FOR TCP & CBR CONNECTIONS

Protocols	Packet Delivery Ratio		Avg. End to End Delay		Loss Packet Ratio	
	TCP	CBR	TCP	CBR	TCP	CBR
AODV	High	High	High	Low	Avg	Low
DSR	High	High	High	Low	Low	Low

TABLE-4: PDR, E-2-E AND LPR WITH RESPECT TO LOW MOBILITY & LOW SPEED TIME FOR TCP & CBR CONNECTIONS

Protocols	Packet Delivery Ratio		Avg. End to End Delay		Loss Packet Ratio	
	TCP	CBR	TCP	CBR	TCP	CBR
AODV	High	High	High	Low	Low	Low
DSR	High	High	High	Low	Low	Low

TABLE-5: PDR, E-2-E AND LPR WITH RESPECT TO LOW MOBILITY & HIGH SPEED TIME FOR TCP & CBR CONNECTIONS

Protocols	Packet Delivery Ratio		Avg. End to End Delay		Loss Packet Ratio	
	TCP	CBR	TCP	CBR	TCP	CBR
AODV	Avg	High	High	Avg	Avg	Avg
DSR	High	Avg	High	Avg	Low	Avg

TABLE-6: PDR, E-2-E AND LPR WITH RESPECT TO HIGH MOBILITY & LOW PAUSE TIME FOR TCP & CBR CONNECTIONS

Protocols	Packet Delivery Ratio		Avg. End to End Delay		Loss Packet Ratio	
	TCP	CBR	TCP	CBR	TCP	CBR
AODV	High	Avg	Avg	Avg	Avg	High
DSR	High	Low	High	High	Avg	High

TABLE-7: PDR E-2-E AND LPR WITH RESPECT TO HIGH MOBILITY & HIGH PAUSE TIME FOR TCP & CBR CONNECTIONS

Protocols	Packet Delivery Ratio		Avg. End to End Delay		Loss Packet Ratio	
	TCP	CBR	TCP	CBR	TCP	CBR
AODV	Avg	High	Avg	Low	High	Low
DSR	Avg	High	High	Low	High	Low

TABLE-8: PDR, E-2-E AND LPR WITH RESPECT TO HIGH MOBILITY & LOW SPEED TIME FOR TCP & CBR CONNECTIONS

Protocols	Packet Delivery Ratio		Avg. End to End Delay		Loss Packet Ratio	
	TCP	CBR	TCP	CBR	TCP	CBR
AODV	Avg	Avg	High	High	High	High
DSR	High	High	High	Low	Low	Low

TABLE-9: PDR, E-2-E AND LPR WITH RESPECT TO HIGH MOBILITY & HIGH SPEED TIME FOR TCP & CBR CONNECTIONS

Protocols	Packet Delivery Ratio		Avg. End to End Delay		Loss Packet Ratio	
	TCP	CBR	TCP	CBR	TCP	CBR
AODV	Avg	Low	High	High	High	High
DSR	High	Low	High	High	Avg	High

## 8.CONCLUSION

In the research paper we mainly analysis the performance of two on demand routing protocols AODV and DSR on the basis of packet delivery ratio, average End-to-End delay and Loss packet ratio. We observe that the performance of AODV and DSR depends on scenario. The performance measurement of AODV and DSR will help for further development of these protocols in future.

## References

- [1] Ericson, "Communication and Mobility by Cellular Advanced Radio", ComCar project, www.comcar.de, 2002.
- [2] <http://www.ist-drive.org/index2.html>.
- [3] W. Franz, H. Hartenstein, and M. Mauve, Eds., *Inter-Vehicle-Communications Based on Ad Hoc Networking Principles-The Fleet Net Project*. Karlsruhe, Germany: Universitatverlag Karlsruhe, November 2005.
- [4] A. Festag, et. al., "NoW-Network on Wheels: Project Objectives, Technology and Achievements", Proceedings of 6th International Workshop on Intelligent Transportations (WIT), Hamburg, Germany, March 2008.
- [5] Reichardt D., Miglietta M., Moretti L., Morsink P., and Schulz W., "CarTALK 2000 - safe and comfortable driving based upon inter-vehicle-communication," in Proc. IEEE IV'02.
- [6] Morris R., Jannotti J., Kaashoek F., Li J., Decouto D., "CarNet: A scalable ad hoc wireless network system," 9th ACM SIGOPS European Workshop, Kolding, Denmark, Sept. 2000.
- [7] Rajesh Deshmukh; Asha Ambhaikar; (*Volume 11- No.8, December 2010* by IJCA.) "Performance Evaluation of AODV and DSR with Reference to Network Size".
- [8] SunilTaneja, AshwaniKush and Amandeep Makkar, (International Journal of Innovation, Management and Technology, Vol. 1, No. 5, December 2010) "Experimental Analysis of DSR, AODV using Speed and Pause time".
- [9] Muazzam Ali Khan Khattak, Khalid Iqbal, Prof Dr. Sikandar Hayat Khoyal (Journal of Theoretical and Applied Information Technology) "Challenging Ad-Hoc Networks under Reliable & Unreliable Transport with Variable Node Density"
- [10] Bijan Paul, Md. Ibrahim, Md. Abu Naser Bikas, "Experimental Analysis of AODV & DSR over TCP & CBR Connections with Varying Speed and Node Density in VANET", International Journal of Computer Applications (IJCA), DOI: 10.5120/2937-3897, Volume 24, Issue 4, June 2011, Pages 30-37
- [11] Bijan Paul, Md. Ibrahim, Md. Abu Naser Bikas, "Performance Evaluation of Aodv&DSR with Varying Pause Time&Node Density Over TCP&CBR Connections in Vanet", International Journal of Computer Science and Network Security (IJCSNS), Volume 11, Issue 7, July 2011, Pages 119-127
- [12] [http://en.wikipedia.org/wiki/List\\_of\\_ad\\_hoc\\_routing\\_protocols](http://en.wikipedia.org/wiki/List_of_ad_hoc_routing_protocols)

- [13] Bijan Paul, Md. Ibrahim, Md. Abu Naser Bikas, "VANET Routing Protocols: Pros and Cons", International Journal of Computer Applications (IJCA), DOI: 10.5120/2413-3224, Volume 20, Issue 3, April 2011, Pages 28-34
- [14] Perkins, C.; Belding-Royer, E.; Das, S. (July 2003)"Ad hoc On-Demand Distance Vector (AODV) Routing".
- [15] Johnson, D. B. and Maltz, D. A. (1996), "Dynamic Source Routing in Ad Hoc Wireless Networks," Mobile Computing, T. Imielinski and H. Korth, Eds., Ch. 5, Kluwer, 1996, pp. 153-81.
- [16] <http://www.isi.edu/nsnam/ns/>
- [17] <http://www.isi.edu/nsnam/ns/tutorial/>

**Bijan Paul** is a B.Sc. student in the Dept. of Computer Science & Engineering, Shahjalal University of Science & Technology, Bangladesh. His research interests include VANET, Routing protocols, Wireless computing.

**Md. Ibrahim** is a B. Sc. student in the Dept. of Computer Science & Engineering, Shahjalal University of Science & Technology, Bangladesh. His research interest includes VANET, Routing protocols, Wireless Computing.

**Md. Abu Naser Bikas** obtained his B. Sc. Degree in Computer Science & Engineering from Shahjalal University of Science & Technology, Bangladesh. Currently, he is a Lecturer in Computer Science & Engineering at the same University. His research interests include VANET, Network Security, Intrusion Detection and Intrusion Prevention, Bangla OCR, Wireless Ad-Hoc Networks, and Grid Computing.

# Assessment of Water Quality in Coastal Environments of Mohammedia Applying Responses of Biochemical Biomarkers in the Brown Mussel *Perna perna*

Laila El Jourmi\*, Abdessamad Amine\*, Meryem Mrani Alaoui, Said Lazar, Abdelaziz Hmyene and Said El Antri

\*These authors contributed equally to this work

Laboratory of Biochemistry Environment and Agroalimentary, Biology Department, University Hassan II, Faculty of Sciences and Techniques Mohammedia, Morocco

## Abstract

The present work aims to assess the marine environment quality in Mohammedia, using the response of the biochemical biomarkers in the brown mussel *Perna perna*. The biomarkers selected in this work are : glutathione S-transferase (GST) as phase II enzyme and the acetylcholinesterase (AChE) activity as neurotoxicity marker. The Oxidative stress is evaluated using catalase (CAT), a well-known anti-oxidant enzyme, and malondialdehyde (MDA) accumulation as marker of oxidation of membrane phospholipids through lipid peroxidation. And finally the metallothioneine (MT) as stress proteins.

Our data indicated that CAT, GST activity and MDA, MT concentration in whole mussel bodies, are a higher and significant ( $p < 0.05$ ) in mussels collected at polluted site when compared to specimen sampled from control one. In contrary the response of AChE activity was significantly ( $p < 0.05$ ) inhibited in mussels from polluted site when compared to control value. The multi-marker results confirm that mussels from Mohammedia have been submitted to polluted environment.

**Keywords:** Biomarkers; Catalase; Malondialdehyde; Glutathione S-transferase; Acetylcholinesterase; Metallothioneine; Mussels; *Perna perna*; Marine Pollution.

## 1. Introduction

Bivalves, in particular mussels, are widely used in biomonitoring programs, mainly due to their biological characteristics; they are sessile, filter-feeding, widely distributed and abundant in coastal and estuarine areas, able to accumulate several classes of pollutants, thus providing a time-integrated picture of their bioavailability. For such characteristics, these organisms are widely used in Mussel Watch monitoring programs, in which chemical analyses are integrated with the use of biomarkers, to evaluate molecular, biochemical and cellular effects induced by pollutants [1].

The measurement of the biological effects of chemical pollutants has become of major importance for the assessment of the quality of the coastal environment. It has

been reported that the use of biomarkers is very informative about the organism's stress response to individual toxicants and mixtures [2-4]. Measuring the same biomarkers in different localities simultaneously gives us information about the pollution status and provides a better comprehension of the mechanistic mode of action of environmental pollutants.

Among these biomarkers are CAT, a well-known antioxidant enzyme, which converts H<sub>2</sub>O<sub>2</sub> into water. The biological importance of CAT is more evident from various studies due to the fact that H<sub>2</sub>O<sub>2</sub> is the main cellular precursor of the hydroxyl radical (HO<sup>-</sup>) which is a highly reactive and toxic form of ROS (Reactive oxygen species) leading to oxidative damage to basic biological molecules.

Toxicity biomarkers, such as MDA, well-known lipid peroxidation products, have been also proposed to reflect the oxidative status of exposed species [5]. MDA is used as marker of oxidation of membrane phospholipids through lipid peroxidation [6]. GST which is a phase II enzyme involved in the metabolism of lipophilic organic contaminants. GST catalyzes the conjugation of various electrophilic compounds (e.g. epoxides of PAHs) with the tripeptide glutathione, the resulting conjugates being water soluble and thus more easily excretable.

AChE is an enzyme essential to the correct transmission of nerve impulses. Its inhibition is directly linked with the mechanisms of toxic action of anticholinesterase compounds [7-9].

MT are useful metal-pollution biomarkers [10], it constitutes a family of low molecular weight, cysteine-rich, and metal binding proteins that occurs throughout the animal kingdom. Biological functions of metallothioneins include homeostasis and sequestration of both essential and nonessential metals, detoxification of metals and scavenging of free radicals [11,10].

The present work aims to assess the marine environment quality in Mohammedia, using the response of the biochemical biomarkers in the brown mussel *Perna perna*.

## 2. Material and Methods

### 2.1 Reagents

Hydrogen peroxide (H<sub>2</sub>O<sub>2</sub>), Thiobarbituric Acid (TBA), Acetylthiocholine (AtChI) and Tetramethoxypropane (TMP) were obtained from Sigma (Saint Quentin Fallavier, France). 1-chloro-2,4-dinitrobenzene (CDNB), 5,5'-dithio-bis 2 nitrobenzoic acid (DTNB), Reduced Glutathione (GSH), and Bovine serum albumin (BSA) were purchased from Genome Biotechnologies (Casablanca, Morocco).

### 2.2 Studied Areas

For this study, two stations (Fig. 1) are selected attending to various degree of human impact. Site 1 (S1) constitutes the site furthest away from the polluting industrial activities established on the coastal fringe Casablanca-Mohammedia. The clean reference area, which is selected due to the absence of contamination sources, located in beach of the south, Skhirat which is characterised by a total prohibition of human activities. S1 is characterized by an important density of mussels and the high faunistic and floristic richnesses of the site are well marked. On the contrary, Site 2 (S2) is located approximately 7 km in South of Mohammedia beach. Due to intense human activities, the S2 is constantly subjected to contamination. In addition, S2 has low biodiversity of intertidal organisms, indicating high levels of pollution pressure.



Fig 1. A map showing mussel sampling locations.

### 2.3 Sampling

Ten mussels from each site were sampled during May 2010. Following collection, the adult mussels were placed in thermally insulated boxes previously filled with water from the sampling site and immediately transported to the laboratory and stored at -80 °C until analysis.

### 2.4 Biochemical Analyses

Whole soft tissues from each specimen (n=5 for each station) were dissected out and immediately homogenized (1:3) in phosphate buffer 100 mM, pH 7.4. Homogenates were then centrifuged at 9000×g at 4 °C for 30 min. After centrifugation, supernatants were collected and immediately used for the determination of enzymatic activity, MDA and MT concentration.

CAT activity was measured following the decrease of absorbance at 240 nm due to H<sub>2</sub>O<sub>2</sub> consumption [12]. The reaction takes place in 100 mM phosphate buffer, pH 7.4 containing 500 mM H<sub>2</sub>O<sub>2</sub>. GST activity was assayed by the method described by Habig et al. (1974) [13] using the CDNB as substrate, and 1 mM GSH, in 100 mM sodium phosphate buffer, pH 7.4. GST activity was determined by kinetic measurement at 340 nm. AChE activity was determined according to the method described by Ellman et al. (1961) [14] using 8 mM DTNB, and 45 mM AtChI as substrate in 100 mM sodium phosphate buffer, pH 7.4. AChE activity was determined by kinetic measurement at 412 nm. Results of enzymatic activities were expressed as nmol /min/mg proteins. MDA was estimated according to the method described by Sunderman (1985) [15] with use of TMP as a standard. The reaction was determined at 532 nm, using TBA as reagent. MDA content was expressed as nmol /mg proteins.

MT content was evaluated in whole soft tissues according to a spectrophotometric method described by Viarengo et al. (1997) [16]. Tissues (n=5 for each station) were homogenised (1:3) in Tris Buffer (Tris 20 mM, 0.5 M sucrose, pH 8.6) containing 0.5 mM phenylmethylsulphonyl fluoride and 0.01% β-mercaptoethanol. The soluble fractions containing MT were obtained by centrifuging the homogenate at 10 000 g for 30 min. The supernatant was then treated with cold absolute ethanol and chloroform. Finally, MT content was spectrophotometrically determined at 412 nm using Ellman's reagent (DTNB) and GSH as standard. The results were expressed as μg/mg proteins. Protein concentrations were measured according to the Bradford (1976) [17] method, at 595 using BSA as standard.

### 2.5 Statistical Analyses

The results for biomarker measurements were investigated by the use of a parametric one-way analysis of variance (ANOVA) and level of significance was set at p < 0.05.

## 3. Results

The marine quality in Mohammedia, is checked in May 2010 at two sites (Fig. 1), by analyzing a battery of biochemical biomarkers of pollution in whole mussel



bodies. For each biomarker, the mean of measurements obtained for each site studied is shown in Fig. 2.

The results relative to the CAT activity are reported in Fig. 2.A. Our results showed that CAT activity presented a higher and significant differences ( $p < 0.05$ ) between polluted and non-polluted stations. The highest CAT activity was registered in mussels from polluted area (S2) with value up to  $22,356 \text{ nmol min}^{-1} \text{ mg}^{-1} \text{ protein}$ , and the lowest activity was reported in mussels from the clean one (S1), with  $12,312 \text{ nmol min}^{-1} \text{ mg}^{-1} \text{ protein}$ .

The response of the phase II enzyme GST activity is shown in Fig. 2.B. A significant differences ( $p < 0.05$ ) values of GST activity were found in mussels collected at the studied stations. It can be seen from Fig. 2.B that The phase II enzyme GST activity values recorded in mussels were significantly lower in animals from S1 ( $23,932 \text{ nmol min}^{-1} \text{ mg}^{-1} \text{ protein}$ ), while those from S2, which is considered as a polluted station, presented the higher significant values ( $28,918 \text{ nmol min}^{-1} \text{ mg}^{-1} \text{ protein}$ ).

The response of AChE activity is reported in Fig. 2.C. The results indicate that AChE activity was significantly ( $p < 0.05$ ) inhibited in mussels from S2 with a value reaching to  $4,01 \text{ nmol min}^{-1} \text{ mg}^{-1} \text{ protein}$  when compared to control value ( $7,284 \text{ nmol min}^{-1} \text{ mg}^{-1} \text{ protein}$ ).

As shown in Fig. 2.D, a higher and significant ( $p < 0.05$ ) accumulation of MDA was registered in *Perna perna* collected at S2 with an amount of  $4,838 \text{ nmole mg}^{-1} \text{ proteins}$  when compared to specimen sampled from control site ( $2,234 \text{ nmole mg}^{-1} \text{ proteins}$ ).

Our data indicated that the levels of MT showed a higher and significant increase ( $P < 0.05$ ) in mussels from S2 ( $4,682 \text{ } \mu\text{g mg}^{-1} \text{ proteins}$ ) in comparison to S1 ( $2,19 \text{ } \mu\text{g mg}^{-1} \text{ proteins}$ ), selected as a suitable reference site (Fig. 2.E).

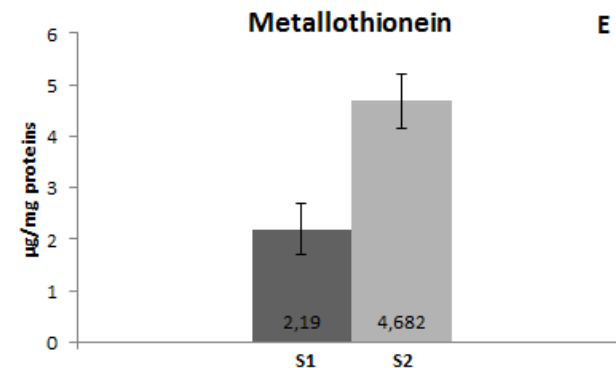
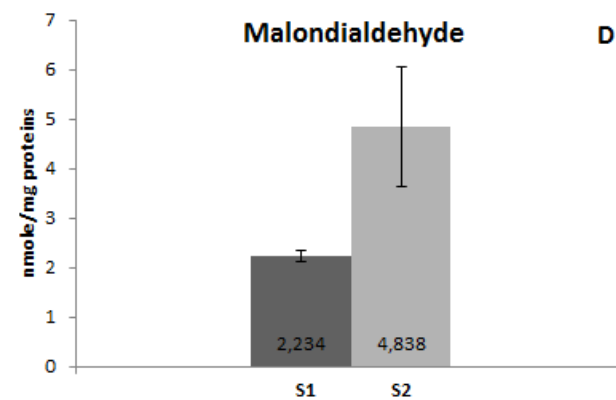
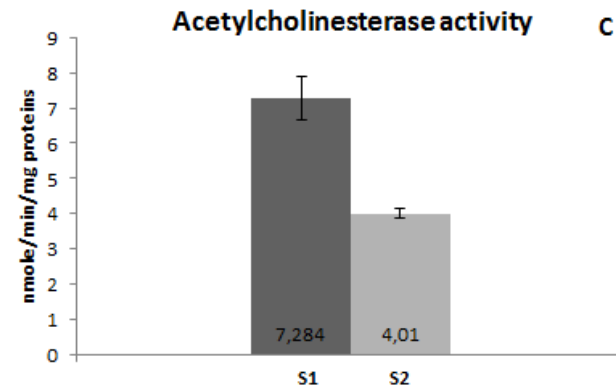
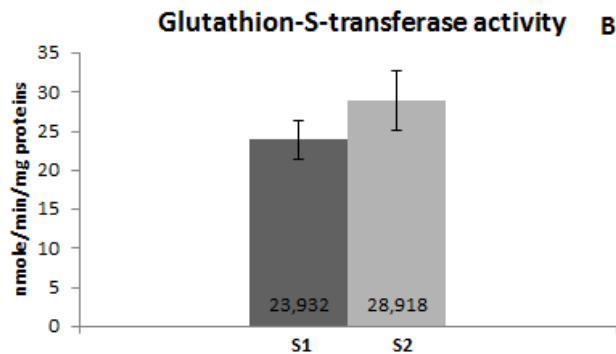
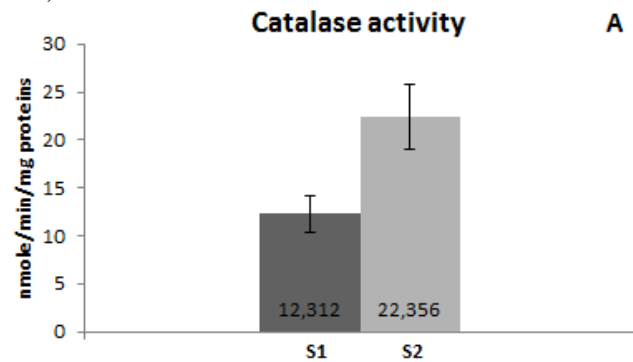


Fig 2. Activities of CAT (A), GST (B), AChE (C), and the levels of MDA (D) and MT (E) in *Perna perna* collected from the studied areas (S1 and S2).

## 4. Discussion

Marine bivalve, especially the mollusks because of their sedentary habitat, are exposed to a variety of environmental stressors, such as heavy metals, pesticides, petroleum products and thermal releases especially in area where cluster of industries are located.

Under normal physiological condition, animals maintain a balance between generation and neutralization of ROS. However when organisms are subjected to xenobiotic compounds, rate of production of ROS, exceeds the scavenging capacity [18] of antioxidant defence systems comprising both antioxidant enzymes, such as CAT, and small molecular weight free radical scavengers.

Exposure of organisms to pro-oxidant attack can increase antioxidant defences by increasing synthesis of antioxidant enzymes [18,19] to counteract oxidative stress. If antioxidant defences are effective in detoxifying ROS, then no harmful consequence results to the tissues. However, if the ROS attack is severe, then antioxidant defence systems may be overwhelmed. Such processes may in turn provoke alterations in molecular and membrane structures and functions leading to cell and tissue damage [20]. The increased activity of CAT in mussels from the investigated site compared to control value indicates that these animals are facing an oxidative challenge, associated with the presence of contaminants in the environment. But, these induced antioxidant defence enzyme increases were not enough to reduce lipid peroxidation levels in the polluted sites. Although CAT removes most of the H<sub>2</sub>O<sub>2</sub> by increasing its activity levels, it cannot compete with the excess presence of H<sub>2</sub>O<sub>2</sub> which diffuses into the cells, thereby causing increased concentration of MDA in S2. According to Charissou et al., (2004) [21], an increase in MDA levels in organisms can be related to degradation of an environmental site by decreasing the water quality. Increased concentration of MDA was observed in mussels exposed in polluted areas when compared to less polluted sites [22,23].

Upon exposure to pollutants, organisms usually attempt to metabolize and depurate directly, minimizing any cellular damages they cause. Such protective mechanisms often include the intervention of enzymes such as GST, a family of multi-functional enzymes [24] involved in Phase II of biotransformation. GST catalyzes the conjugation of various electrophilic compounds (e.g. epoxides of PAHs) with the tripeptide glutathione, the resulting conjugates being water soluble and thus more easily excretable. The toxicity of many exogenous compounds can be modulated by induction of GST. Indeed GST activities were found to be modulated by metals or organic contaminants under both field conditions [25,28] and laboratory exposure [29]. This study demonstrated that mussels collected at site S2

presented significant higher levels of GST activity than those collected at site S1. So their induction can be regarded as an adaptive response to an altered environment. Several other field studies have demonstrated a similar relationship between environmental contamination and GST activity in mussels [30,31].

The main physiological function of AChE is splitting of acetylcholine, a mediator of cholinergic synapses, during transduction of nerve impulses [32]. The inhibition of AChE activity has been widely used to diagnose exposure to anticholinesterase compounds [7-9]. Many studies indicate that Cholinesterase activities are inhibited in the presence of some pesticides [33,34]. In fact, in addition to anticholinesterase pesticides, a few other contaminants such as heavy metals, detergents, some pyrethroids compounds like cypermethrin and deltamethrin and complex mixtures of pollutants can also affect the AChE activity [35,39]. In our study, we demonstrate that mussels inhabiting the site 2 (S2) present a higher inhibition of AChE activities compared to the site 1 (S1). The observed inhibition of AChE activities may be attributed to the presence of contaminants in the environment.

MT constitutes a family of low molecular weight, cysteine-rich, and metal binding proteins that occurs throughout the animal kingdom. Although the expression of MTs may be influenced by a number of natural factors [40] such as water temperature, they are also clearly induced by some of the environmentally most important heavy metals, making MT a potential biomarker for metal exposure in the aquatic environment [41]. Indeed the induction of MTs as a measure of response to metal exposure in aquatic organisms has been widely investigated in laboratory and field conditions [42-44].

It has been reported that MTs are known to be involved in heavy metal homeostasis and overexpressed in organisms experiencing high metal concentrations in their environment [45,46]. In this work, we have found that MT accumulation level is higher in mussels from S2 in comparison to S1. Concerning the role of MT in the homeostasis of trace metals and the elevated concentrations of these elements in mussels from S2, might compensate the metal bioavailability. So the elevated MT level at the site S2 may be considered as the result of higher overall level of metals pollution.

## 5. Conclusions

In conclusion, our data confirm the pollution status of site S2, which subjected to various types of pollution sources. The main sources of pollution are industrial, due to the presence of S2 in industrial park, and urban activities. Thus agricultural activities developed in the area, generate important quantities of toxic substances able to

contaminate the surface and deep waters whose general flow is directed towards the Ocean. So the geographical distribution of site 2 can explain the alteration of oxidative parameters.

## References

- [1] D.R., Livingstone. Biotechnology and pollution monitoring: use of molecular biomarker in the aquatic environment. *J. Chem. Technol. Biotechnol.*, 1993, 57, 195–211.
- [2] C., Svendsen, D.J., Spurgeon, P.K., Hankard, J.M., Weeks. A review of lysosomal membrane stability measured by neutral red retention: is it a workable earthworm biomarker? *Ecotoxicol. Environ. Saf.*, 2004, 54, 20–29.
- [3] L., Gastaldi, E., Ranzato, F., Capri, P., Hankard, G., Pérès, L., Canesi, A., Viarengo, G., Pons. Application of a biomarker battery for the evaluation of the sublethal effects of pollutants in the earthworm *Eisenia Andrei*. *Comp. Biochem. Physiol Part C*, 2007, 146, 398–405.
- [4] O., Adam, P.M., Badot, F., Degiorgi, G., Crini. Mixture toxicity assessment of wood preservative pesticides in the freshwater amphipod *Gammarus pulex* (L.). *Ecotoxicol. Environ. Saf.*, 2009, 72, 441–449.
- [5] M., Sole, C., Porte, X., Biosca, C.L., Mitchelmore, J.K., Chipman, D.R., Livingstone, J., Albaiges, Effects of the “Aegean Sea” oil spill on biotransformation enzymes, oxidative stress and DNA-Adducts in digestive gland of the mussel (*Mytilus edulis* L.). *Comp. Biochem. Physiol. Part C Pharmacol. Toxicol.*, 1996, 113, 257–265.
- [6] M.L., Alexandrova, P.G., Bochev. Oxidative stress during the chronic phase of stroke. *Free Radical Biol. Med.*, 2005, 39, 297–316.
- [7] G., Rodriguez-Fuentes, & G., Gold-Bouchot. Characterization of cholinesterase activity from different tissues of Nile tilapia (*Oreochromis niloticus*). *Marine Environmental Research*, 2004, 58, 505–509.
- [8] E., Ozcan-Oruc, & D., Usta. Evaluation of oxidative stress responses and neurotoxicity potential of diazinon in different tissues of *Cyprinus carpio*. *Environmental Toxicology and Pharmacology*, 2007, 23, 48–55.
- [9] F. R., De la Torre, L., Ferrari, & A., Salibian. Freshwater pollution biomarker: Response of brain acetylcholinesterase activity in two fish species. *Comparative Biochemistry and Physiology Part C*, 2002, 131, 271–280.
- [10] A., Viarengo, B., Burlando, N., Ceratto, I., Panfoli. Antioxidant role of metallothioneins: a comparative overview. *Cellular and Molecular Biology*, 2000, 46, 40–417.
- [11] I., Bremner, J.H., Beattie. Metallothionein and the trace minerals. *Annual Review of Nutrition*, 1990, 10, 63–83.
- [12] H., Aebi. Catalase. In: Bergmeyer, H.U. (Ed.), *Methods of Enzymatic Analysis*, vol. 3. Academic Press, New York, 1983, pp. 237–286.
- [13] W. H., Habig, M. J., Pabst, & W. B., Jakoby. Glutathione S-transferases, the first step in mercapturic acid formation. *The Journal of Biological Chemistry*, 1974, 249(22), 7130–7139.
- [14] G.L., Ellman, D., Courtneyk, V., Andres, R.M., Featherstone. A new and rapid colorimetric determination of acetylcholinesterase activity. *Biochem. Pharmacol.*, 1961, 7, 88–95.
- [15] F.W., Sunderman. *Annals of clinical and laboratory science.*, 1985, 13 (3) ; 229.
- [16] A., Viarengo, E., Ponzano, F., Dondero, & R., Fabbri. A simple spectrophotometric method for metallothionein evaluation in marine organisms: An application to Mediterranean and Antarctic molluscs. *Marine Environmental Research*, 1997, 44, 69–84.
- [17] M., Bradford. A rapid and sensitive method for the quantification of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical Biochemistry*, 1976, 72, 248–254.
- [18] B., Halliwell and J.M.C., Gutteridge. *Free Radicals in Biology and Medicine*. Oxford Univ. Pres, Oxford. Hodgson EK Fridovich I. (1975): *The Interaction of Bovine Erythrocyte Superoxide Dismutase with Hydrogen Peroxide: Inactivation of the Enzyme*. *Biochemistry*, 1999, 14, 5294–5298.
- [19] M.J., Bebianno, R., Company, A., Serafim, R.P., Cosson, A., Fiala-Medoni. Antioxidant systems and lipid peroxidation in *Bathymodiolus azoricus* from Mid-Atlantic Ridge hydrothermal vent fields. *Aquat. Toxicol*, 2005, 75, 354–373.
- [20] A.Z., Reznick, L. Packer and C.K., Sen. Strategies to Assess Oxidative Stress. In: A.Z., Reznick, L., Packer, C.K., Sen, J.O., Holloszy and M.J., Jackson (eds.): *Oxidative Stress in Skeletal Muscle*. Birkhauser Verlag, Basel, Switzerland, 1998, 43–48.
- [21] A.M., Charissou, C., Cossu-Leguille, P., Vasseur. Relationship between two oxidative stress biomarkers, malondialdehyde and 8-oxo-7,8-dihydro- 2'-deoxyguanosine, in the freshwater bivalve *Unio tumidus*. *Sci. Tot. Environ*, 2004, 322, 109–122.
- [22] P.S., Lau, H.L., Wong. Effect of size, tissue parts and location on six biochemical markers in the green-lipped mussel, *Perna viridis*. *Mar. Pollut. Bull.*, 2003, 46, 1563–1572.
- [23] D.M., Pampanin, L., Camus, A., Gomiero, I., Marangon, E., Volpato, C., Nasci. Susceptibility to oxidative stress of mussels (*Mytilus galloprovincialis*) in the Venice Lagoon (Italy). *Mar. Pollut. Bull.*, 2005, 50, 1548–1557.
- [24] A., Pathirante, & S. G., George. Oxidative malathion to Nile tilapia, *Oreochromis niloticus* and modulation by other environmental contaminants. *Aquatic Toxicology*, 1998, 43, 261–271.
- [25] J. A., Knight, R. K., Pieper, & L., McClellan. Specificity of the thiobarbituric acid reaction: Its use in studies of lipid peroxidation. *Clinical Chemistry*, 1988, 34, 2433–2438.
- [26] P. J., Fitzpatrick, J., O'Halloran, D., Sheehan, & A. R., Walsh. Assessment of a glutathione-S-transferase and related proteins in the gill and digestive gland of *Mytilus edulis* (L.) as potential organic pollution biomarkers. *Biomarkers*, 1997, 2, 51–56.
- [27] P., Hoareau, M., Gnassia-Barelli, M., Roméo, & J. P., Girard. Differential induction of GST in the clam *Ruditapes decussatus*. *European Journal of Biochemistry*, 2001, 269, 4359–4366.
- [28] C., Durou, L., Poirier, J.C., Amiard, H., Budzinski, M., Gnassia-Barelli, K., Lemenach, L., Peluhet, C., Mouneyrac, M., Romeo, C., Amiard-Triquet. Biomonitoring in a clean and a multi-contaminated estuary based on biomarkers and

- chemical analyses in the endobenthic worm *Nereis diversicolor*. *Environ. Pollut.*, 2007, 148, 445–458.
- [29] D. R., Livingstone. Organic xenobiotic metabolism in marine invertebrates. *Advanced Comparative Environmental and Physiology*, 1991, 7, 145–187.
- [30] C.C.C., Cheung, G.J., Zheng, A.M.Y., Li, B.J., Richardson, P.K.S., Lam. Relationships between tissue concentrations of polycyclic aromatic hydrocarbons and antioxidative responses of marine mussels, *Perna viridis*. *Aquat. Toxicol.*, 2001, 52, 189–203.
- [31] S.M., Moreira, M., Moreira dos Santos, R., Ribeiro, L., Guilhermino. The ‘Coral Bulker’ fuel oil spill on the North coast of Portugal: spatial and temporal biomarker responses in *Mytilus galloprovincialis*. *Ecotoxicology*, 2004, 13, 619–630.
- [32] C. H., Walker, & H. M., Thompson. Phylogenetic distribution of cholinesterases and related esterases. In P. Mineau (Ed.), *Cholinesterase-inhibiting insecticides. Their impact on wildlife and the environment*. Amsterdam: Elsevier, 1991, 3–19.
- [33] F., Galgani, & G., Bocquené. A method for routine detection of organophosphates and carbamates in sea water. *Environmental Technology Letters*, 1989, 10, 311–322.
- [34] P., Scaps, & O., Borot. Acetylcholinesterase activity of the polychaete *Nereis diversicolor*: Effects of temperature and salinity. *Comparative Biochemistry and Physiology, (Part C)*, 2000, 125, 377–383.
- [35] T., Balint, T., Szegletes, & H., Szegletes. Biochemical and subcellular changes in carp exposed to the organophosphorus methidathion and pyrethroid deltamethrin. *Aquatic Toxicology*, 1995, 33, 279–295.
- [36] L., Guilhermino, M. N., Lacerda, A. J. A., Nogueira, & A. M. V. M., Soares. In vitro and in vivo inhibition of *Daphnia magna* acetylcholinesterase by surfactant agents: Possible implications for contamination biomonitoring. *Science of the Total Environment*, 2000, 247, 137–141.
- [37] M. P., Cajaraville, M. J., Bebianno, J., Blasco, C., Porte, C., Sarasquete, & A., Viarengo. The use of biomarkers to assess the impact of pollution in coastal environments of the Iberian Peninsula: A practical approach. *Science of the Total Environment*, 2000, 247(2–3), 295–311.
- [38] M., Banni, J., Jebali, M., Daubeze, C., Clerandau, H., Guerbej, J. F., Narbonne. Monitoring pollution in Tunisian coasts: Application of a classification scale based on biochemical markers. *Biomarkers*, 2006, 10(2–3), 105–116.
- [39] J., Jebali, M., Banni, E. A., Almeida, A., Bannaoui, & H., Boussetta. Effects of malathion and cadmium on acetylcholinesterase activity and metallothionein levels in the fish *Seriola dumerilli*. *Fish Physiology and Biochemistry*, 2006, 32, 93–98.
- [40] D., Ivankovic, J., Pavicic, M., Erk, V., Filipovic-Marjic, B., Raspor. Evaluation of the *Mytilus galloprovincialis* Lam. digestive gland metallothionein as a biomarker in a long-term field study: seasonal and spatial variability. *Mar. Pollut. Bull.*, 2005, 50, 1303–1313.
- [41] G., De Boeck, T. T., Huong Ngo, K., Van Campenhout, & R., Blust. Differential metallothionein induction patterns in three freshwater fish during sublethal copper exposure. *Aquatic Toxicology*, 2003, 65, 413–424.
- [42] H., Reynders, L., Bervoets, M., Gelders, W. M., De Coen, & R., Blust. Accumulation and effects of metals in caged carp and resident roach along a metal pollution gradient. *Science of the Total Environment*, 2008, 391, 182–195.
- [43] H., De Smet, B., De Wachter, R., Lobinski, & R., Blust. Dynamics of (Cd, Zn)-metallothioneins in gills, liver and kidney of common carp *Cyprinus carpio* during cadmium exposure. *Aquatic Toxicology*, 2001, 52, 269–281.
- [44] A., Viarengo, D., Lowe, C., Bolognesi, E., Fabbri, & A., Koehler. The use of biomarkers in biomonitoring: A 2-tier approach assessing the level of pollutant-induced stress syndrome in sentinel organisms. *Comparative Biochemistry and Physiology Part C*, 2007, 146, 281–300.
- [45] A. Viarengo. Heavy metals in marine invertebrates, mechanisms of regulation and toxicity at cell level. *CRC Critical Review in Aquatic Science*, 1989, 1, 295–317.
- [46] A., Viarengo, B., Burlando, M., Cavaletto, B., Marchi, E., Ponsano, & J., Blasco. Role of metallothionein against oxidative stress in the mussel *Mytilus galloprovincialis*. *American Physiological Society*, 1999, 363, 1614–1617.

Please consider to contribute to and/or forward to the appropriate groups the following opportunity to submit and publish original scientific results.

## CALL FOR PAPERS International Journal of Computer Science Issues (IJCSI) Volume 9, Issue 3 – May 2012 Issue

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.

Indexed by Google Scholar, DBLP, CiteSeerX, Directory for Open Access Journal (DOAJ), Bielefeld Academic Search Engine (BASE), SCIRUS, Cornell University Library, ScientificCommons, EBSCO, ProQuest and more.

**Deadline: 31<sup>st</sup> March 2012**

**Notification: 30<sup>th</sup> April 2012**

**Revision: 10<sup>th</sup> May 2012**

**Publication: 31<sup>st</sup> May 2012**

Context-aware systems  
Networking technologies  
Security in network, systems, and applications  
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  
Knowledge for global defense  
Information Systems [IS]  
IPv6 Today - Technology and deployment  
Modeling  
Software Engineering  
Optimization  
Complexity  
Natural Language Processing  
Speech Synthesis  
Data Mining

For more topics, please see <http://www.ijcsi.org/call-for-papers.php>

arXiv.org

Google scholar

SCIRUS  
search engine for science

ScientificCommons

Scribd

.docstoc  
find and share professional documents



BASE  
Bielefeld Academic Search Engine

CiteSeer<sup>beta</sup>

dblp.uni-trier.de  
Computer Science  
Bibliography

DOAJ  
DIRECTORY OF  
OPEN ACCESS  
JOURNALS

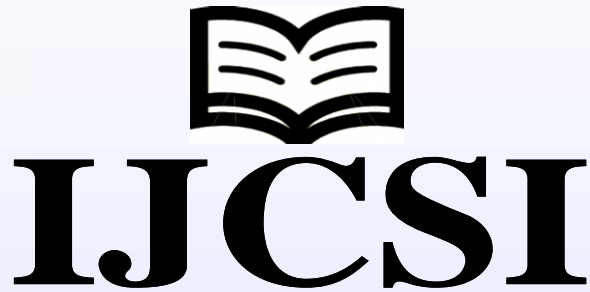


ProQuest

For more information, please visit the journal website ([www.IJCSI.org](http://www.IJCSI.org))

**© IJCSI PUBLICATION 2012**

**[www.IJCSI.org](http://www.IJCSI.org)**



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.

#### **Indexing of IJCSI**

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