An Efficient Method for Urdu Language Text Search in Image Based Urdu Text

Khalil Khan [1], Muhammad Siddique [2] , Muhammad Aamir[3] & Rehanullah Khan [4]

[1] Department of Computer System Engineering, University of Engineering & Technology, Peshawar, 25000, Pakistan

[2] Department of Computer System Engineering, University of Engineering & Technology, Peshawar, 25000, Pakistan

[3] Department of Electrical Engineering, Sarhad University of Science & Information Technology, Peshawar, 25000, Pakistan

[4] Department of Electrical Engineering, Sarhad University of Science & Information Technology, Peshawar, 25000, Pakistan

Abstract

This paper describes an efficient method for Urdu text search in computer generated and handwritten scanned images. An efficient text search technology is necessary because of increasing handled document every day. This method is unique and simple in the sense that no features are extracted. The proposed method is script independent. The input image is directly matched with a set of prototype characters representing each possible class. The distance between each input image and each prototype character is computed, and the character is assigned to the class of the prototype giving the best match. Experimental results show 100 % accuracy for 4, 5-character ligatures, 87 % for 3-character ligature and 78 % for 2-character ligatures.

Keywords: *Template matching, correlation analysis, optical character recognition, source image, template image.*

1. Introduction

Urdu is the national language Pakistan and is spoken in more than 20 countries of the world [1]. Speakers of Urdu are between 60 and 70 million. Persian, Arabic and Turkish have great influence on Urdu language and this is the reason Urdu is a mixture of all these languages. Its writing style is from right to left. Arabic and Farsi languages have close resemblance with Urdu, but Urdu is more complex as compare to Arabic and Farsi due to additional characters. Therefore recognition methods of Arabic and Farsi are not applicable to Urdu. Urdu OCR is still unsolved problem and research is still in progress in Urdu OCR. Therefore rather than using complex methods of Urdu OCR an easy method of text search in Urdu image based text is introduced.

Personal computer is spreading rapidly and general people are using electronic documents such as email, newspapers, books etc written in Urdu in large amount. Reading Urdu newspapers and other stuff on internet and computer are common nowadays as it is a time saving and cheap way. Almost all newspapers like daily jang, daily Nawai Waqt, daily ajj, daily Khabrain and many more are available on internet and are written using Urdu Inpage. Urdu language support is also available now for windows XP, windows vista, Linux, MS office etc. Many online libraries are available on internet having handwritten scanned books and computer generated books as well. Urdu along with English is the official language of Pakistan and is used in courts, educational system, offices, literature etc. All these show that Urdu text is increasing day by day and a high speed and efficient text search technology is necessary for handling these documents.

Optical character recognition is the conversion of the text images into machine editable text. This technology is very developed for other languages like English, German, Chinese, Arabic etc. But Urdu optical character recognition is not so much developed due to two reasons; complexities involved in the characters of Urdu language and lack of centralized system which fund the research work of Urdu language. Hence we do not have any complete OCR system which we can apply for text search in Urdu image based text.

Remaining paper is organized as follows: Section 2 describes related work, Section 3 explains template matching based approach, section 4 proposed methodology, Section 5 filtering text based Image, Section 6 describes matching in text based images, Section 7 experimental results and Section 8 concludes.

2. Previous Work

Research work on Urdu optical character recognition is still in progress. U. Pal and Anirban Sarkar et al [2] conducted study on single character recognition. This technique involves Skew detection and correction, line segmentation and Character segmentation. Complex and

compound characters or ligatures cannot be recognized by this technique. Inam Shamsher et al [3] proposed a technique for recognition of single characters only claiming 98.3 % accuracy. This method is script independent and is for printed characters only. No details have been given of the tool or software used in the method. Zaheer Ahmad et al [4] published a paper on Urdu Optical character recognition for Nastalique font. Segmentation is divided into three steps according to the authors. Lines of the text are identified in the first step. Words are separated in second step and final step show characters segmentation and extraction from words. Data obtained in previous step is then given to neural network for classification and recognition. 93.4 % accuracy is claimed by the authors of the paper without giving experimental details. S. A. Husain et al [5] proposed a method for recognition of single character, 2 character and 3 character ligatures. Segmentation free approach has been used in this paper. After pre processing stage feature vector is extracted for ligature and then passed to the BPNN for classification purpose. This method is for Nastaliq script. Tabassam Nawaz et al [6] presented an idea for character recognition of isolated characters. The Method proposed by them works basically in three steps; image pre processing, segmentation of line and character, making Xml file which is then used as a database and for training purposes. The authors claim 89 % recognition accuracy. Sobia Tariq Javed et al [7] worked on pre processing stage only. Their work is for Nastalique style only. In first stage the horizontal line or base line is detected and separated. In next stage ligature base and then diacritics are segmented. According to the authors of the paper 100 % accuracy has been reported for base line identification and 94 % accuracy for ligature identification.

We have avoided using typical methods of Urdu character recognition due to some reasons; the current OCR methods for Urdu language still need lots of research work; most of the OCR methods are for single characters only as table 1 show and Urdu language has very few words with single characters; methods which are applicable to ligature is script dependent [7] and our proposed method is script independent; template matching approach is fast and robust as compare to typical methods of Urdu OCR. Hence we introduced a fast and easily applicable method for text search in Urdu based image.

Table 1: Current OCR methods.

Authors	Specification	Accuracy
U. Pal and Anirban Sarkar et al	Isolated characters	96.9 %
Inam Shamsher et al	Isolated characters	98.3 %
Zaheer Ahmad et al	For ligature	93.4 %
S. A. Husain et al	For ligature	93 %
Tabassam Nawaz et al	Isolated characters	89 %

3. Template Matching Approach

Template matching approach has various applications. Some of these applications are face detection [8], visual object recognition [9], car plate recognition [10], Human Ear Detection[11], On road vehicle detection[12] etc. Nadira Muda et al [13] applied template matching approach for English alphabets recognition as well. Mohammed Ali Qatran [14] used template matching approach for Musnad character recognition. In this paper we present an idea how to find a ligature or character inside a larger document. Template matching is a simple method used for classification and pattern recognition. Individual image pixels are used as feature in case of template matching. The template image is compared to each template of the source image to find an exact match or a template with nearest representation in database. A database of templates is used for matching in this process which is also called training set. This method can identify scanned or computer written characters, numbers and other secondary characters also called diacritics. In template matching, the template image is moved to all possible positions of the source image and exact match or a match with nearest representation is found. All matching process is pixel-by-pixel basis.

If I(x,y) is the input template image, T(x,y) is the source image in the database, O(x,y) the output image, M(x,y) will give an exact match or nearest match of the input character or ligature.

$$M(c,T) = \sum_{i=0}^{w}\sum_{j=0}^{h}|c(i,j) - T(i,j)| \quad (1)$$

$$M(c,T) = \sum_{i=0}^{w}\sum_{j=0}^{h}(c(i,j) - T(i,j))^2 \quad (2)$$

$$M(c,T) = \sum_{i=0}^{w}\sum_{j=0}^{h}(c(i,j)T(i,j))^2 \quad (3)$$

Equation (1) shows city block, equation (2) shows Euclidean distance and equation (3) shows cross correlation.

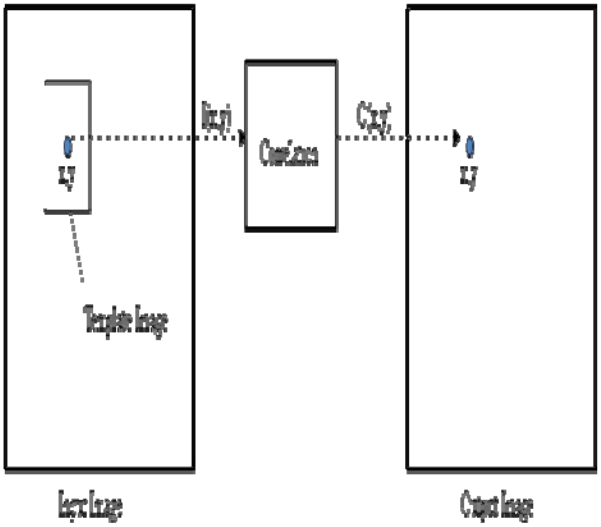


Figure 2: Template matching approach

4. Proposed Methodology

The proposed approach consists of the following steps:

- Reading of handwritten Scanned image or computer generated image.
- Filtering the text based image for Noise removal.
- Conversion of image from RGB to grayscale if necessary.
- Image Recognition using template matching and correlation algorithm.
- Conversion of Recognized characters into text and if there is no match then displaying the result of 'No match found'.

Block diagram in figure (1) illustrates the proposed methodology.

5. Filtering Text Based Image:

Image contains wanted and unwanted data. Image filtering is the process of removal of unwanted data. Usually computer written images have no noise but scanned images contain noise such as speckle noise and salt and pepper noise. Several types of filters are used for noise removal e.g minimum filtering, maximum filtering, median filtering, average filtering etc. Median filtering technique has been used in the proposed method. Median filter has property that high frequency details are not lost while removing noise. Thus edge blurring not occurs while performing filtering. This filter removes salt and pepper noise and Gaussian noise as well. Matlab function medfilt2(image) has been used for this purpose.

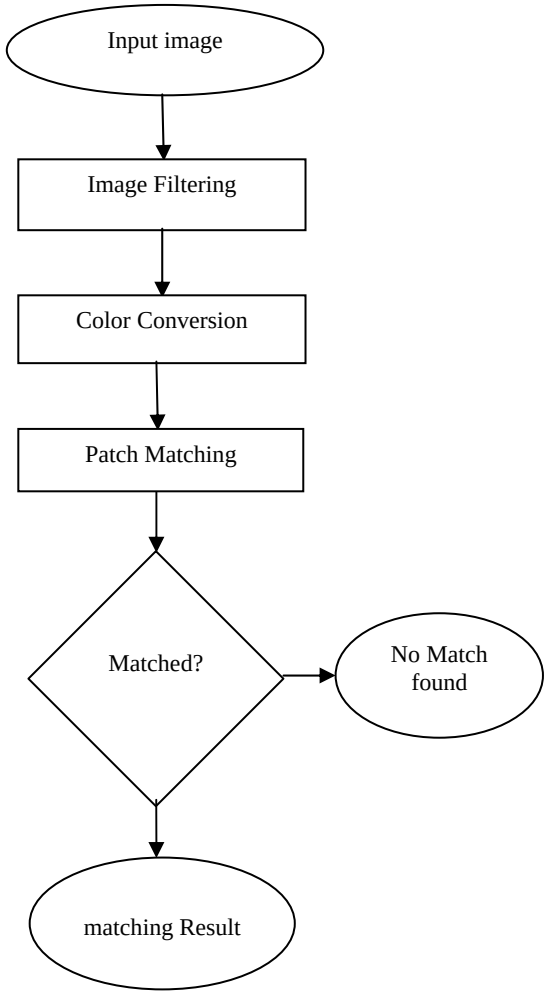


Figure 1: The proposed methodology for Urdu text search.

6. Matching

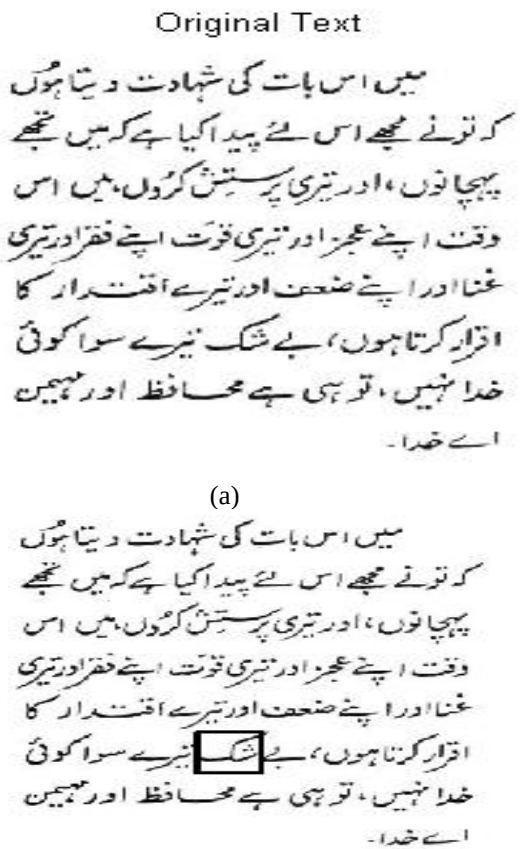
Urdu characters and ligatures are extracted by first removing noise from images. Dimensions of the source image were kept 512 x 512 and that of the template image 42 x 24. The source images were stored in the database. The template image is matched with all the images in the database using 2-D correlation matching function in Matlab. During matching process a numerical index is computed. This numerical index shows how the template matches the source image in that specific position. Template matching approach has some limitations; templates are not rotation invariant, dimension reducing can cause problems and it is computationally an expensive process.

7. Experimental Results:

All the experiments were performed using Matlab as software tool. 200 test images were taken including handwritten and computer generated images. These test images were single line and many lines as well. Initially resolution of image was kept high. Results recognition was quite good for high resolution images. However, as we dropped resolution of images then recognition rate was lowered. It was also noted during experiments that recognition rate of ligature having more characters was higher. Recognition rate was 100 % for four, five character ligature, 87 % for three characters ligature and 78 % for two character ligature. Experimental results are shown in figure (2).

Table 2: Recognition rate for 2, 3, 4 and 5-character ligatures.

Image Resolution	Ligature	Accuracy
512 x 512, 42 x 24	2-characters	78 %
512 x 512, 42 x 24	3-characters	87 %
512 x 512, 42 x 24	4-characters	100 %
512 x 512, 42 x 24	5-characters	100 %



(a)

(b)

(c)

(d)

(e)

Figure 2: Characters Search: (a) Original image, (b) 2-characters (c) 3-characters (d) 4-characters, and (e) 5-characters search.

8. Conclusions and Future Work

A simple and robust method of finding a character or ligature in Urdu text images is introduced in this paper. An Urdu text image is scanned or written using computer. Noise is removed from image and then converted into grayscale image. A set of images were taken including Urdu characters and ligatures. The image to be searched is taken as a template image and then it is compared with the source image which is used as a database. The template image or the image with closest representation is found in the source image. 2-D correlation coefficient approach has

been used between the test image and the database source image.

Results show that template matching approach can be used to find a character or whole ligature inside an image easily. This method can be then expanded for a complete Urdu OCR system. Template matching approach can be combined with character recognition using HMM and neural networks as well. In future we will try to develop a complete OCR system for Urdu recognition using the approach used in the paper.

References

[1] Raymond G. Gordon, "Ethnologue: languages of the World Fifteenth Edition" SIL International, 2005.
[2] U. Pal and Anirban Sarkar "Recognition of Printed Urdu Script", Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003), IEEE.
[3] Inam Shamsher, Zaheer Ahmad, Jehanzeb Khan Orakzai, and Awais Adnan "OCR For Printed Urdu Script Using Feed Forward Neural Network", World Academy of Science, Engineering and Technology 34 2007.
[4] Zaheer Ahmad, Jehanzeb Khan Orakzai, Inam Shamsher, and Awais Adnan "Urdu Nastaleeq optical character recognition", World Academy of Science, Engineering and Technology 32 2007
[5] S. A. Husain, Asma Sajjad, Fareeha Anwar "Online Urdu Character Recognition System", MVA2007 IAPR Conference on Machine Vision Applications, May 16-18, 2007, Tokyo, JAPAN.
[6] Tabassam Nawaz, Syed Ammar Hassan Shah Naqvi, Habib ur Rehman, Anoshia Faiz "Optical Character Recognition System for Urdu (Naskh Font) Using Pattern Matching Technique", International Journal of Image Processing, (IJIP)Volume (3) : Issue (3).
[7] Sobia Tariq Javed and Sarmad Hussain, "Improving Nastalique-Specific Pre-Recognition Process for Urdu OCR", Multitopic Conference, 2009. INMIC 2009. IEEE 13th International.
[8] Smita Tripathi, Varsha Sharma and Sanjeev Sharma,"Face Detection using Combined Skin Color Detector and Template Matching Method", International Journal of Computer Applications (0975 – 8887) Volume 26– No.7, July 2011.
[9] Luke Cole, David Austin, Lance Cole, "Visual Object Recognition using Template Matching", Proceedings of Australian Conference on Robotics and Automation, 2004.
[10] M.I.Khalil, "Car Plate Recognition Using the Template Matching Method", International Journal of Computer Theory and Engineering, Vol. 2, No. 5, October, 2010 1793-8201
[11] K. V. Joshi, N. C. Chauhan, "Edge Detection and Template Matching Approaches for Human Ear Detection", International Conference on Intelligent Systems and Data Processing (ICISD) 2011 Special Issue published by International Journal of Computer Applications® (IJCA)
[12] Rajiv Kumar Nath, Dr. Swapan Kumar Deb, "On road vehicle/object detection and tracking using template", Indian Journal of Computer Science and Engineering Vol 1 No 2, 98-107
[13] Nadira Muda, Nik Kamariah Nik Ismail, Siti Azami Abu Bakar, Jasni Mohamad Zain, Fakulti Sistem Komputer & Kejuruteraan Perisian, "Optical Character Recognition By Using Template Matching (Alphabet)", National Conference on Software Engineering & Computer Systems 2007 (NACES 2007).
[14] Mohammed Ali Qatran "template matching method for recognition Musnad characters based on correlation analysis", ACIT'2011 Proceedings.

Khalil Khan is pursuing his M.Sc. computer System engineering from University of Engineering & Technology, Peshawar, Pakistan. He has completed his B.Sc. Electrical engineering from the same university. He worked as switching engineer at ZTE pvt. ltd. in Optical fiber and Switching department for two and half year. He also worked as a lecturer in Electrical engineering department, University of Lahore, Islamabad campus. Currently he is working as a Principal Dir College of Science & Technology. His research interest areas are image processing, computer vision, machine learning and pattern recognition.

Muhammad Sidique did B.Sc. Computer System engineering from University of engineering & Technology, Peshawar, Pakistan. He is also pursuing his M.Sc. computer System engineering from University of Engineering & Technology, Peshawar. He is working as a System Engineer in Ghulam Ishaq Khan Institute of Science & Technology, Topi, Pakistan

Muhammad Aamir
Muhammad Aamir completed his B.E and M.SC from UET peshawar. Currently, he is working as an Assistant Professor at Sarhad University of Science and IT, Peshawar, Pakistan.

Dr. Rehanullah graduated from University of Engineering and Technology Peshawar, with a BSc degree (Computer Engineering) in 2004 and MSc (Information Systems) in 2006. He obtained PhD degree (Computer Science) in 2011 from Vienna University of Technology, Austria. He is currently an Assistant Professor at Sarhad University of Science and Technology, Peshawar. His research interests include color interpretation, segmentation and object recognition.