Urdu Character Recognition using Principal Component Analysis

Khalil Khan
Department of Computer System Engineering,
University of Engineering and Technology, Peshawar, Pakistan.

Rehan Ullah
Department of Electrical Engineering,
Sarhad University of Science and Information Technology, Peshawar, Pakistan.

Nasir Ahmad Khan
Department of Computer System Engineering,
University of Engineering and Technology, Peshawar, Pakistan.

Khwaja Naveed
Department of Electrical Engineering,
University of Engineering and Technology, Peshawar, Pakistan.

ABSTRACT
This paper proposes a method for Urdu language text search in image based Urdu Text. In the proposed method two databases of images have been created; first one for training purpose and another for testing purpose. Training database is named 'TrainDatabase' and testing database as 'TestDatabase'. Training database consists of all characters of Urdu language in different shapes. Eigen values and Eigen vectors of all the images to be placed in the TrainingDatabase are calculated. Only those values having highest Eigen values are kept. A feature vector for each image of the TrainDatabase is calculated by the algorithm. A threshold value is chosen such that it defines maximum allowable distance between TrainDatabase and TestDatabase images. Feature vector is also created for each image to be identified and placed in 'TestDatabase'. Comparison is done for a character to be identified with each image of 'TrainDatabase'. If the character to be recognized is matching with any character of the TrainDatabase result is shown by algorithm. MATLAB has been used as a simulation tool and the recognition rate obtained was 96.2 % for isolated characters.

General Terms
Pattern Recognition, Optical Character Recognition.

Keywords
Optical Character Recognition, Principal Component Analysis, Training Database, Testing Database.

1. INTRODUCTION
Optical character recognition is an active area of research. It is the conversion of hand written and machine written text to computer readable/ editable form. Researchers have proposed several methods for recognition of different languages such as English, Chinese, Arabic [1, 2, 3] etc. Commercial OCR for most of these languages is also developed and available in market. However, for Urdu language research is still in initial stages to the best of my knowledge and lots of research is needed in this field.

Speakers of Urdu language are over 60 millions [4]. National language authority of Pakistan has defined 58 character set for Urdu language. However only 40 of these characters are used in Urdu literature. Character set for Urdu language is shown in Figure 1.

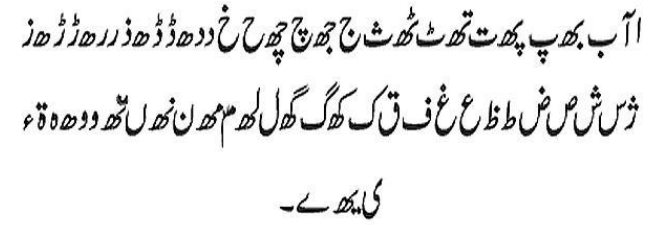

Figure 1: Character Set for Urdu Language

Urdu character recognition is a challenging task due complexities involved in Urdu language. These complexities are described bellow;

Ligature: Several characters are combined to form a ligature [5]. The ligature consists of more than two characters combined to form a single word. Figure 2 shows how characters are combined to form a single word.

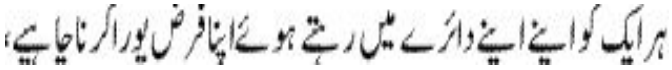

Figure 2: Characters Connectivity Problem

Diacritics: Urdu words are combination of primary and secondary characters. These secondary characters are called diacritics. Diacritics include dots, hamzas, diagonal etc. These secondary characters are placed bellow or above the primary characters. Figure 3 shows various diacritics in various positions.

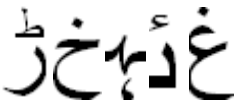

Figure 3: Various Diacritics in different Positions

Sensitivity of context: Each character in Urdu has 2-4 different shapes. These shapes depend on position such as beginning, middle, and in last. Some characters also have isolated shapes. Figure 4 shows various positions of character Ein.

Figure 4: Isolated, Starting, Ending and middle Character

Writing Direction: Languages spoken in the world are uni directional in maximum cases. But Urdu language is unique in this regard. It is bidirectional language.

Base Line: It is a horizontal line which is present in the whole text .

Scripts: Urdu language has 12 different scripts. Most of the recognition methods are script specific and does not work on recognition of other scripts.

All the above difficulties make recognition of Urdu characters a challenging task.

Remaining Paper is arranged in this order; section 2 explains literature review of the work done in recognition of Urdu language; section 3 describes theory of PCA; section 4 explains mathematics of PCA; section 5 is about the method used in the paper; section 6 tell details of the experimental results and section 7 is about the conclusion and future work to be performed in OCR of Urdu language.

2. Literature Review

Image based text search is an important area of research of image processing and pattern recognition. U. Pal and Anirban Sarkar et al [6] worked was related to recognition of isolated characters. The whole method is divided into three steps. During first step of the process skew detection and correction of a character is performed. In next stage of the process base line and character segmentation is performed. Recognition phase is performed during last stage of the whole process. According to the authors of the paper the method was tested on both noisy images and images without noise. The authors claim 98.3% accuracy for base line recognition while the accuracy for individual characters and numerals was 97.8%. Inam Shamsher et al [7] proposed a technique for Urdu character recognition. This method is also for individual characters recognition. The authors of the paper used feed forward neural network. The neural network has been trained by using supervised learning. According to the authors of the paper the proposed method is script independent. The proposed method achieves 98.3% accuracy according to the authors of the paper. No features have been extracted from characters of the image and individual image pixels are used for learning purpose. Due to this factor accuracy of the proposed method is low. Zaheer Ahmad et al [8] published a paper on Nastalique font. The proposed method is divided into three steps. These steps are Base line identification, words separation and character segmentation and features extraction. The data obtained is then given to neural network for training. After training stage classification and then recognition is performed in last step. This method is also for a specific font Nastalique of Urdu language. According to the authors of the paper accuracy of the proposed method is 93.4%. S. A. Husain et al [9] published a paper for recognition of single, double and triple character ligature. Unlike other traditional methods of recognition no segmentation is performed in this method. First step performed is Pre processing after which a feature vector is created. BPNN has been trained by this feature vector which is used in later stages for classification and recognition. This method is also for a specific font and works only on Nastalique font. Tabassam Nawaz et al [10] also published paper on isolated character recognition. First step of the recognition is image pre processing. During which noise is removed from the image if it contains. Line and character segmentation is performed in next stages. The data obtained in all these steps is then used for training purposes during which an xml file is created. During recognition stage the image to be recognized is segmented. Chain code of the character to be recognized is made and then it is matched with the xml file which is already created and works as database. According to the authors of the paper recognition rate of the proposed method is 89% and speed of recognition is 15 character/second. Sobia Tariq Javed et al [11] work was limited only to pre processing stage and Nastalique font only. The proposed work is only part of some improvements to the existed methods. Base line is separated in first stage of the process. In next stages segmentation of ligature and diacritics is performed. Authors claimed 100 % accuracy for baseline identification and 94 % for ligature identification.

3. Principal Component Analysis

PCA is widely used in Image Processing, Pattern Recognition and Computer Vision. It is basically a statistical method purpose of which is dimensionality reduction and interpretation of data [12]. Charles Spearman in 1904 published a paper on principal component analysis in American journal of psychology for the first time. He applied PCA to the data set of social sciences. He was studying human intellectual ability while taking various matters such as critical reasoning, writing, mathematics and sciences. Based on his analysis using PCA he introduced an intelligence factor which was named later on as IQ factor.

Processing of large data set is not only difficult but almost impossible. Main aim of PCA is dimensionality reduction and features extraction. By reducing dimensionality of data, speed of computation is increased enormously without losing important information. First step of Principal component analysis is training of data. A set of images is taken for training purposes initially. PCA transform of the training data is calculated which is actually computation of Eigen vectors and Eigen images. Training data is projected onto Eigen space. After the training stage testing data is also projected in the same way. As a last stage classification of the data is done while comparing training and testing data.

Application area of PCA is very vast. Most important application of Principal component analysis is face recognition [13, 14]. Khaled Labib et. al [15] used principal comment analysis for detection of Network Attacks. Kavita Mahajan et al [16] used PCA for Classification of Electroencephalogram (EEG) in his published work. PCA also has applications in Image Denoising, Intrusion detection. Similarly it is also been used for extraction of features from English characters.

4. Mathematics of Principal Component Analysis

Two dimensional images are converted into one dimensional image by concatenation of rows and columns into a vector. Let we have M vectors having a set of images and let p_j's represent pixel values of the images

$X_i = [p_1, p_2, \ldots\ldots p_N]^T$, i= 1,……,M

The mean of each image is obtained by

$$m = \frac{1}{M} \sum_{t=1}^{M} x_i$$

ω_i which is the mean centered image will be

$\omega_i = X_i - m$

Goal of PCA is calculation of a set of e_i's having maximum projection onto each of ω_i 's. Let we want to calculate a set of M orthonormal vectors for which

$$\lambda_i = \frac{1}{M} \sum_{n=1}^{M} x_{i(e_i^T \omega_n)^2}$$

is maximum having orthonormality constraint

$e_i^T e_K = \sigma_{ik}$

Eigen vectors and values of covariance matrix C gives the values λ_i's and e_i's and the covariance matrix can be written as bellows

$C = WW^T$

If we represent Eigen vectors and Eigen values by d_i and $\mu_{i\ of}$ WW^T, then

$WW^T d_i = \mu_i d_i$

Multiplying both hand sides by W

$WW^T (Wd_i) = \mu_i (Wd_i)$

Which means that Eigen vectors and Eigen values of WW^T are Wd_i and μ_i respectively.

The Eigen vectors are stored from highest to lowest value to their corresponding Eigen values. The Eigen vector having maximum Eigen value shows greatest variance in image.

Image of a character to be find is projected onto M' dimensions and the following value is calculated

$\Omega = [v_1, v_2, \ldots, v_{M'}]^T$

Where

Ω shows the contribution of each Eigen Image.

$v_i = e_i^T \omega_n$

The vectors e_i are Eigen images or sometime also called Eigen faces.

The method for determining which character provides the nearest description of an input character is to find the class k that has minimum Euclidean distance. Euclidean distance can be calculated as;

$\epsilon_k = \|\Omega - \Omega_k\|$

Where

$\Omega_k = k^{th}$ Eigen image class

The character or numeral belongs to class k if value of Ω_k is less than a threshold which is predefined. In pattern recognition this threshold is represented by θ_C.

5. Proposed Methodology

Following steps are followed during methodology proposed in the paper.

Step 1: First step of the proposed method is pre processing. During pre processing stage noise is removed from images of TrainDatabase and TestDatabase. Speckle noise and salt and pepper noise is present in the scanned images normally.

Step 2: A database of images named 'TrainDatabase' has been created. This database consist characters of Urdu language. The database has been used for training purposes later on in the algorithm. Training set includes different images (M) of each character with some variation in writing style and shape. Different writing styles are chosen so that there is no problem in classification stage.

Step 3:A second database of images called 'TestDatabase' is also created. The image of the character to be identified is placed in this database.

Step 4:A matrix L (MxM) has been calculated for each image of 'TrainDatabase'. Eigen vectors and Eigen values for this matrix are found. M' Eigen vector is chosen such that it has highest associated Eigen values.

Step 5: Feature vector is created for each image of 'TrainDatabase'. This value is stored which is used for classification. Feature vector is also created for each character to be identified containing in 'TestDatabase'.

Step 6:A threshold value is chosen such that it defines maximum allowable distance for each class. This threshold value is used for classification purposes.

Step 7: Feature vector of a character to be recognized is compared with the stored feature vectors of the 'TrainDatabase'. If the character to be identified is matching with any character of 'TrainDatabase' then the 'Known' character is given by the algorithm. Figure 5 shows detail of proposed methodology

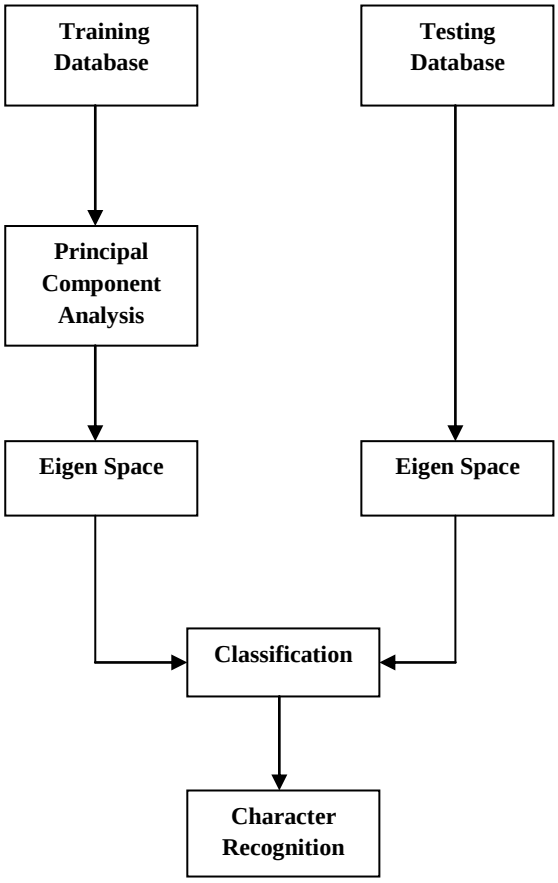

Figure 5: Proposed Methodology for Character Recognition

6. Experimental work and results

MATLAB R2009b has been used as an experimental tool in the proposed work. The TrainDatabase has been used for training and classification purpose. Image to be recognized is placed after pre processing stage in TestDatabase. Characters of Urdu language consisting of both hand written and machine written text were tested during experiments. As each character in Urdu language has 2-4 shapes and maximum shapes of the

images were taken for training purposes. After PCA transform of training images is taken, the training images are projected onto Eigen space. Similarly image to be recognized is also projected onto Eigen space and classification is done. Recognition rate for noise free images was good but as images having noise were taken recognition rate was dropped. The individual character recognition rate of 96.2 % was noted during experiments.

Recognition rate for single characters with no secondary characters was noted to be highest. However when secondary characters were added, accuracy was lowered. These results are shown in table 1.

Table 1: Recognition Rate for different characters

S No.	Character Type	Recognition Rate
1	Single Character without secondary characters	96.2 %
2	Single character with single secondary character	95.4 %
3	Single character with two secondary characters	94.2 %

7. Conclusion and Future Work

Two databases 'TrainDatabase' and 'TestDatabase' have been created in MATLAB for the proposed work. TrainDatabase has been used for training and classification purposes. Characters to be recognized are placed in TestDatabase. PCA transform of all images in TrainDatabase is calculated by the PCA algorithm. Training images are then projected onto Eigen space. Test images which are to be recognized are also projected onto Eigen space. Classification is done by the algorithm and the character to be recognized is shown by the algorithm.

The proposed method works on recognition of isolated characters only. This method can be extended for complete word/ligature recognition as well. The proposed method can be combined with artificial neural network, Support vector machine to make a complete OCR of Urdu language. The feature set obtained through PCA can also be used in Weka tool for efficient classification and recognition.

8. References

[1] Wei Zhao, Jia-Feng Liu ; Xiang-Long Tang "Online handwritten English word recognition based on cascade connection of character HMMs", Machine Learning and Cybernetics, 2002. Proceedings. 2002 International Conference, vol.4 pp. 1758 - 1761

[2] G. Nagy Rensselaer Polytechnic Institute Troy, New York, "Chinese Character Recognition A Twenty Five Year Retrospective". Tsuyoshi Kitani t, riguchi and Masami Ilara Yoshio, "Pattern Matching in the Textract Information Extraction System".

[3] T.S El-Sheikh and R.M Guindi, "computer Recognition of Arabic Cursive Script," Pattern Recognition, Vol.21, No, 4, 1988, pp.293-302.

[4] Raymond G. Gordon, "Ethnologue: Languages of the World Fifteenth Edition" SIL International, 2005.

[5] Zahra A Shah and Farah Saleem. "Ligature Based Optical Character Recognition of Urdu, Nastaleeq Font", INMIC 2002.

[6] U. Pal and Anirban Sarkar "Recognition of Printed Urdu Script", Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003), IEEE.

[7] 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.

[8] Zaheer Ahmad, Jehanzeb Khan Orakzai, Inam Shamsher, and Awais Adnan "Urdu Nastaleeq optical character recognition", World Academy of Science, Engineering and Technology 32 2007

[9] 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.

[10] 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).

[11] Sobia Tariq Javed and Sarmad Hussain, "Improving Nastalique-Specific Pre-Recognition Process for Urdu OCR", Multitopic Conference, 2009. INMIC 2009. IEEE 13th International.

[12] J Edward. A User's Guide to Principal Components. New York: Wiley-Interscience, 1991.

[13] M.A. Turk and A.P. Pentland, "Face Recognition Using Eigenfaces", *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 586-591, 1991.

[14] M.S.R.S. Prasad, S.S. Panda, G. Deepthi and V. Anisha "Face Recognition Using PCA and Feed Forward Neural Networks", International Journal of Computer Science and Telecommunications [Volume 2, Issue 8, November 2011].

[15] Khaled Labib and V. Rao Vemuri , "An Application of Principal Component Analysis to the Detection and Visualization of Computer Network Attacks" Conference on Security and Network Architectures the proceedings of SAR 2004.

[16] Kavita Mahajan, M. R. Vargantwar, Sangita M. Rajput, "Classification of EEG using PCA, ICA and Neural Network", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-1, Issue-1, October 2011