Urdu Text Classification Using Decision Trees

K. Khan*, R. Ullah khan#, Ali Alkhalifah#, N. Ahmad**,
* Dept. of Information Engineering, University of Brescia, Italy
# College of Computer, Qassim University, KSA
** Dept. of Computer Systems Engineering, University of Engineering and Technology Peshawar, Pakistan
k.khalil@unibs.it, n.ahmad@uetpeshawar.edu.pk, rehanmarwat1@gmail.com, alkhalifahali@gmail.com

***Abstract*—this article reports the development and experimental analysis of an Urdu Optical Character Recognition (OCR) system. The proposed approach presents the preprocessing, features extraction and classification of Urdu language text. Three different features extraction techniques, the Hu moments, Zernike moments and the Principal Component Analysis (PCA) are used. Decision Tree algorithm J-48 is used for classification. A medium size database of 441 characters is created consisting of hand written and machine written Urdu language characters. An overall best recognition accuracy of 92.06 % is achieved using the Hu moments.**

Keywords: classification, decision tree algorithm, feature extraction, Urdu optical character recognition.

1. INTRODUCTION

OCR finds its applications in banks, shopping malls, health care centers, railway stations, airports, passport offices etc. As application areas of OCR technology is increasing, new problems surfaces. OCR is still in infancy stages for most of the world languages and a lot of research work is yet needed to make the OCR fulfill the needs of these applications.

Urdu is the national language of Pakistan, also spoken and understood in India and the 5th largest language community of the world. Urdu is a complex language as compared to other languages such as English, Chinese, and German etc. It has close resemblance with Arabic language but its OCR is more complicated than Arabic. In Arabic language total numbers of characters in use are 28 [1], while Urdu language has 58 characters set including basic characters and the derived characters which are made of the combinations of basic characters. Out of these 58 characters, 41 characters are used frequently in Urdu literature [2]. A set of 58 Urdu characters is shown in Figure 1. We believe that the Urdu OCR is a challenging task due to the following reasons:

Ligature: In Urdu language, two or more characters are combined to form a single word which is called ligature [3]. This complex combination is a major challenge for OCR.

Diacritics: Most characters of Urdu language are not single but there are some symbols bellow or above the primary character. These symbols are called secondary characters or diacritics. Recognition of these diacritics produces problems in segmentation and then in recognition stages.

Context sensitivity: Each character in a ligature has different shape depending on the position of the character. Character may occur in start, middle and ending. Isolated shapes of characters also exist. Due to context sensitivity, Urdu language is completely different from other languages.

Direction of writing: Urdu language is bidirectional in nature. Its characters are written from right to left and its numerals are written from left to right.

Diagonal writing style: Words and characters are written in Urdu language from right diagonal for beauty and sufficient space. On one hand, it creates beauty in writing style but on the other hand, it creates overlapping problems.

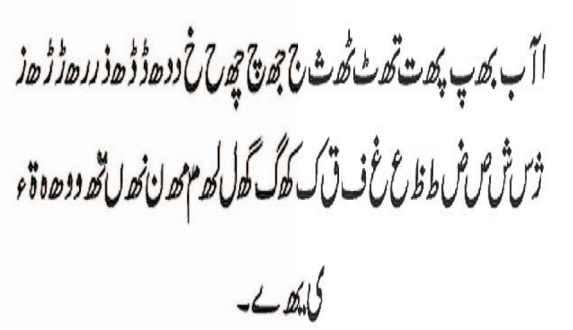

Figure 1: Urdu characters set

The Dotting Problem: Characters and words are differentiated from each other in Urdu language using dots.

Script: Urdu language has 12 different writing scripts. The recognition method may be script dependent or independent. Most of the recognition methods are designed for a specific script.

2. PREVIOUS WORK

Urdu language has some resemblance with the Arabic. Substantial work on Arabic OCR has been done as compared to Urdu OCR [4]. Similarly, Chinese and Japanese OCR are in very developed form [5, 6]. Hussain *et al.* [7] in his work proposed a method which can recognize limited number of ligatures only. Artificial neural network has been used in the proposed method. U Pal *et al.* [8] presented his study about recognition of isolated characters only. They extracted features from Urdu characters using water reservoir feature extraction methods. Shamsheer *et al.* [9] also addressed the recognition of isolated characters. Feed Forward Neural

978-1-4673-9268-6/15/$31.00 ©2015 IEEE

Network has been used for classification purposes. Durrani *et al.* [10] discussed segmentation problems which Urdu OCR is facing. Ahmad *et al.* [2] work is specific for Nastalique Urdu script only. The proposed method works in steps; lines identification; words separation; character segmentation and features extraction; and finally, classification. Classification was performed in the proposed method using Neural Network. Authors of the paper claimed 93.4% accuracy for isolated characters recognition. Nawaz *et al.* [11] proposed a method for isolated characters recognition. The proposed method work in three steps: pre-processing, segmentation and classification. For classification purpose, the authors presented a new idea of an XML file creation. This XML file is used as a database and source for classification. Authors of the paper claimed 89 % accuracy for isolated character recognition. Javed *et al.* [12] proposed idea for the improvement of preprocessing stage only. The proposed method is organized as; Detection and separation of Horizontal Base Line; Ligature base segmentation and Diacritics segmentation. Base line is separated and identified accurately. Accuracy for Ligature identification is 94%.

3. Decision Tree Algorithm

Decision tree (DT) is a classification algorithm that generates decision tree based on the training data. The generated decision tree data is also called classification data. The classification is done on the bases of 'divide and conquer' strategy [13]. Structure of decision tree is hierarchical in nature where a test is applied at each level to the attribute values having one of two outcomes. For classification, root of the tree is taken as base, test is evaluated, and a branch is chosen with appropriate outcome. The above process continues until a leaf is reached and it is decided whether it belongs to the class named by a specific leaf or not. Each leaf is generated as a result of a set of mutually exclusive decision rules, as we move along the tree. The tree expands until each and every instance is classified correctly or incorrectly. J-48 is a decision tree algorithm used in the proposed work. The decision tree is generated using the concept of information entropy. The J-48 algorithm incorporates information gain ratio as attribute splitting criterion [14, 15].

4. Features

Features capture most important information from the characters. This information is then passed for classification and recognition. Various features extraction techniques have been devoloped by researcher in text recognition. In the proposed work, we use three different methods for features extraction; Hu Moments, Zernike Moments and features extraction using PCA.

Hu Moments [16] are introduced by Hu. These feature extraction methods extract invariant moment features from images. Hu moments have been extensively used in Pattern recognition and image processing [17]. Hu moments of order 2 to 9 have been used. Experiments with other feature extraction methods indicate that at least 10-15 features are needed to be extracted for a successful and accurate OCR system [18].

Zernike moments have already been used for character recognition of other languages [19]. Zernike moments are a set of orthogonal polynomials over polar coordinates inside a unit circle [20]. Zernike moments calculate Euclidean distance between the character to be recognized and the training image.

PCA is an un-supervised learning technique which is mostly used for features extraction in Pattern Recognition [21]. Similarities and differences in a dataset are identified easily by PCA. PCA is comparatively an easy tool for analyzing dataset having higher dimensions. After finding a specific pattern, dimensions of the data is reduced by a compression process without losing important information.

5. Proposed Methodology

Figure 2 shows the block diagram of the proposed methodology. MATLAB is used for preprocessing and feature extraction, while classification is performed using J-48. As no reference database exists for Urdu characters, a database of Urdu characters consisting of both hand written and machine written scanned images is created. The database consists of 441 Urdu language characters. Various steps for the proposed methodology are;

A. Preprocessing

Preprocessing step reduces processing operations for further steps to be followed. Steps in preprocessing in the proposed methodology are:

Noise Removal: For noise removal in OCR, two main approaches exist; filtering (masks) and morphological operations (Erosion and dilation). Filtering technique has been adapted in the proposed work. Filtering process is performed with Median filter.

Binarization: is performed in character recognition by two approaches; global thresholding and adaptive thresholding. Global thresholding is used in the proposed algorithm.

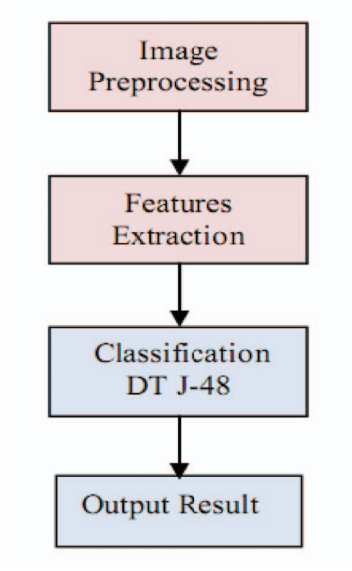

Figure 2: Proposed Methodology

Normalization of the image: Two main objectives are achieved using normalization; data size reduction and thinning. Thinning process extracts only shape information from characters.

Skew Correction: Skew correction is the process of alignment of the paper document with system of the scanner. Correlation approach has been used in the proposed work for achieving skew correction.

Edge detection: Edge detection identities boundaries in an image. Some researcher include edge detection in feature extraction technique, however in the proposed work, we included it in the preprocessing steps.

B. *Features Extraction and Classification*

After preprocessing stage, features are extracted from the characters. The feature set used include the Hu moments, Zernike moments and PCA. Each feature vector is extracted separately from each character and then passed to the classifier. These feature vectors are processed by J-48 for classification and analysis.

6. Results and Discussion

The proposed approach of experimental analysis for OCR has three main steps; Preprocessing; Features extraction and Classification; as depicted in Figure 2. Initial two steps are performed using MATLAB while classification is performed using J-48.

Table 1 and Figure 3 summarize results for correctly classified instances, incorrectly classified instances, time taken to complete the model and the Kappa statistics. Table 2 summarizes the results for the mean absolute error, root mean squared error, relative absolute error and root relative squared error.

Table 1: Performance of the proposed approach

Features Extracted	Correctly classified instances	Incorrectly classified instances	Time Taken	Kappa Statistics
Hu moments	*92.06 %*	*7.93 %*	*0.14 sec*	*0.91*
Zernike moments	*54.32 %*	*45.07 %*	*0.26 sec*	*0.43*
PCA	*32.13 %*	*71.87 %*	*0.46 sec*	*0.32*

Table 2: Various error measures of the proposed model

Features Extracted	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Root Relative Squared Error
Hu moments	*0.004*	*0.06*	*8.16 %*	*40.39 %*
Zernike moments	*0.67*	*0.32*	*26.34 %*	*67.15 %*
PCA	*0.73*	*0.45*	*39.86 %*	*75.74 %*

Accuracy of the proposed model is highest (92.06 %) for Hu moments. Total numbers of instances taken are 441. An average number of instances classified correctly were 406 and incorrectly classified instances are 35. For Zernike moments, the accuracy yield is 54.32 %. Out of 441 instances, 238 instances were classified correctly and 202 incorrectly. Lowest classification rate (32.13 %) is observed for the PCA approach. An average of 142 instances was correctly classified and 298 instances were incorrectly classified. Table 1 shows highest accuracy for Hu moments. Time taken to build a model is also an important parameter in recognition scenarios. When Hu moments are used for feature extraction, the time taken to build a model was 0.14 seconds, which is the fastest in terms of execution times.

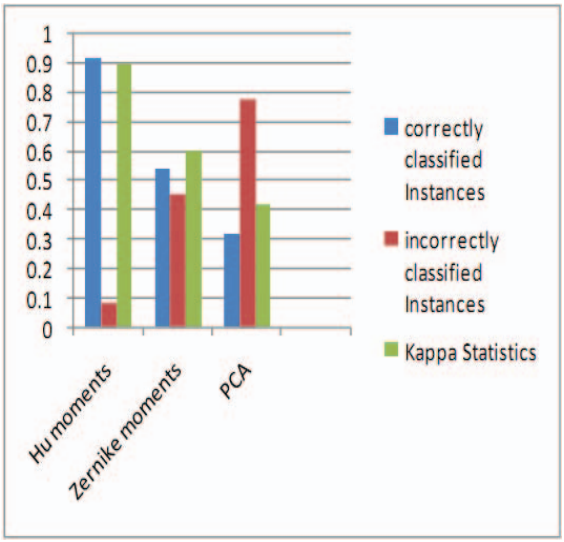

Figure 3: Performance comparison of Hu Moments, Zernike Moments and PCA

Kappa statistic is used to assess the accuracy of any particular measuring cases. It is generally used to distinguish between the reliability of the data collected and their validity. It is analogous to correlation coefficient. The Kappa score ranges from 0.3 to 0.9 during the proposed work. Kappa statistics value for PCA is 0.32 showing a weak statistical model. Kappa statistics value for Zernike moments is 0.43, indicating a moderate statistical dependence. For Hu moments, its value is 0.91 indicating that the model could advantageously be used for recognition purposes. If the value of the Receiver Operating Characteristics (ROC) area is near 0.5, it shows lack of statistical dependence. The weighted Average ROC value for Hu moments is 0.95, for Zernike moments ROC value was 0.73 and for PCA 0.63.

7. Conclusion

This work analyzed Urdu OCR using the three feature extraction approaches and the decision tree classification algorithm. As no reference database for Urdu characters exits, as a first step, a database of Urdu characters is created. An experimentation setup included 441characters. Highest recognition accuracy of 92.06 % is obtained with the Hu moment.

REFERENCES

[1] Majed Ismail Hussien, Fekry Olayah, Minwer AL-dwan & Ahlam Shamsan "Arabic Text Classification Using SMO Navie Bayesian, and J-48 algorithm " International Journal of research and review in applied sciences, volume 9, Issue 2, 2011.

[2] Zaheer Ahmad, Jehanzeb Khan Orakzai, Inam Shamsher, and Awais Adnan " Urdu Nastalique OCR s", Proceedings of World Academy of Science, Engineering and Technology, Volume 2, ISSN:1307-6884, 2007.

[3] Umar Iftikhar, "Recognition of Urdu Ligatures", M.Sc. Thesis German Research Center for Artificial Intelligence, (DFKI) 2011.

[4] Khorsheed, Mohammad S, "Offline Arabic character recognition- a review", Pattern analysis & applications, 5(1):31- 45, 2002.

[5] C.L. Liu, S. Jaeger, and M. Nakagawa. "Online recognition of Chinese characters: the state-of-the-art." Pattern Analysis and Machine Intelligence, IEEE Transactions on, 26(2):198-213, 2004.

[6] Hisamitsu, Toru, et al. "Optimal techniques in OCR error correction for Japanese texts." *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*. Vol. 2. IEEE, 1995.

[7] Husain, Syed Afaq. "A multi-tier holistic approach for Urdu Nastaliq recognition." *Multi topic conference, 2002. Abstracts. INMIC 2002. International*. IEEE, 2002.

[8] U. Pal and A. Sarkar "Recognition of printed Urdu script", Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003).

[9] I. Shamsher, Z. Ahmad, J.K. Orakzai, and A. Adnan "OCR for printed Urdu script using feed forward neural network", Proceedings of World Academy of Science, Engineering and Technology, volume 23. , 2007.

[10] N. Durrani and S. Hussain, "Urdu word segmentation", In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Com-mutational Linguistics, pages 528-536. Association for Computational Linguistics, 2010.

[11] 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).

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

[13] Rasoul Safavian and David Landgrebe, "A survey of Decision Tree Classifier Methodology", IEEE Transactions on Systems, Man and Cybernetics, Volume 21, No.3, May/June 1991.

[14] H. Ian Witten and Eibe Frank, "Data Mining: Practical machine learning tools and techniques", Elsevier, 2nd edition, 2005.

[15] Quinlan, J Ross, "Decision Trees and Decision making", IEEE transactions on Systems, Man and Cybernetics, Vol.20, No.2, March/April 1990.

[16] Hu MK, "Visual Pattern Recognition by Moment Invariants", IRE Transactions on Information theory, IT (8), pp. 179 -187, 1962.

[17] T.H. Resiss, "the revised fundamental theorem of moment invariant", IEEE Trans. Pattern analysis and machine intelligence, volume 13, pp. 830-834, Aug 1991.

[18] Oivind Due Trier, Anil K.Jain, and Torfinn Taxt, "Feature extraction method for character recognition- a survey", Pattern Recognition, volume 29, No.4, pp. 641-662,1996.

[19] Mukundan R. and Rmakrishnan K. R. "Moments functions in image analysis theory and applications", world scientific publishing, Singapore, 1998.

[20] Otiniano-Rodríguez, K. C., G. Cámara-Chávez, and D. Menotti. "Hu and Zernike moments for sign language recognition." Proceedings of international conference on image processing, computer vision, and pattern recognition. 2012.

[21] M.A. Turk and A.P. Pentland, "Face Recognition Using Eigenfaces", IEEE Conf. on Computer Vision and Pattern Recognition, pp. 586-591, 1