Multi-font Numerals Recognition for Urdu Script based Languages

Muhammad Imran Razzak, S.A. Hussain, Abdel Belaïd, Muhammad Sher

Multi-font Numerals Recognition for Urdu Script based Languages

Muhammad Imran Razzak[1], S.A.Hussain[2], [3]A. Belaid and Muhammad Sher[1]
[1] International Islamic University, Islamabad, Pakistan
Email: {imranrazak@hotmail.com, msher@iiu.edu.pk}
[2]Air University, Islamabad, Pakistan.

***Abstract*—Handwritten character recognition of Urdu script based languages is one of the most difficult task due to complexities of the script. Urdu script based languages has not received much attestation even this script is used more than 1/6th of the population. The complexities in the script makes more complicated the recognition process. The problem in handwritten numeral recognition is the shape similarity between handwritten numerals and dual style for Urdu. This paper presents a fuzzy rule base, HMM and Hybrid approaches for the recognition of numerals both Urdu and Arabic in unconstrained environment from both online and offline domain for online input. Basically offline domain is used for preprocessing i.e normalization, slant normalization. The proposed system is tested and provides accuracy of 97.1% .**

***Index Terms*—Numerals Recognition, Online Urdu Numerals, Handwriting, Urdu**

I. Introduction

Character recognition is the process of converting the language written in spatial form into computer understandable form (Unicode). Online Recognition provides natural ways to interact with machine without having any typing skills. The character recognition is classified into two main domains with respect to input namely online and offline (Offline is further divided into two categories printed and handwritten). In offline image is available while in online strokes are available with timing information. Thus due to the additional timing information online character recognition is easy than handwritten offline character recognition.

The main problem of online Urdu handwritten digit recognition is to extract proper feature matrix, because both Urdu and Arabic are written. The use of digitizing tablet/light pen makes the data entry easy, flexible and its is a natural way of input while it is difficult to afford keyboard for enormous entries.

Numeral for Urdu is written in Arabic script but commonly Roman numeral is used. Urdu, Farsi and Hindi numerals look very similar with minor difference [1]. Indian (Hindi, Urdu) digits are used mostly in Arabic countries while Farsi digits are used mainly in Iran. The Indian digits are normally written in 11 classes. Figure [1] compares the Arabic (used in English and Latin), Farsi and Indian (Hindi). Now a day the mostly written digits are written in old Arabic figure 1 digit forms. Urdu numerals consist of line segments and curves and different arrangement of these numerals form different numbers as shown in figure 1 and figure 2.

Urdu digits mostly followed in Pakistan are shown in figure 2. Generally, in Urdu both Urdu numerals and old Arabic (used for English, Latin) are followed. Urdu is written in both but not at the same time. Basically numerals for Urdu are same like Farsi script but commonly written in Old Arabic numeral instead of Urdu numerals.

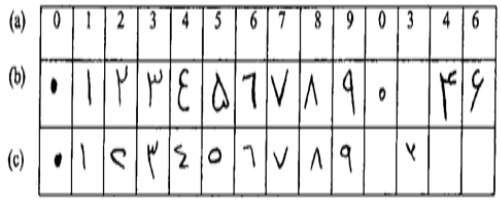

Figure. 1. (a) Arabic digits (used mainly in Latin and English countries but also in others), (b) Farsi digits (used in Iran), (c) Indian (Hindi) digits (used in Arabic countries).

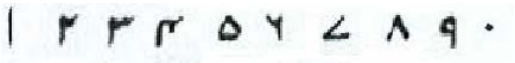

Figure. 2. (a) Urdu digits (used mostly in subcontinent),

A. Harifi et.al [2] presented a technique for Persian digits using multilayer preceptiron and proposed asymmetrical segmentation pattern for feature extraction and 12 segment was used and used the shadow coding and 97.6% recognition is reported. S.V. Rajashekararadhya [4] presented zone centroid and image centroid based distance metric feature extraction system for Indian script numeral recognition. The numerals centroid is computed and the numeral image is divided in to n equal zones. Average distance from centroid to the each pixel in the zone is computed. Nearest neighbor and feed forwards back propagation is used for classification and 99 %, 99%, 96% and 95 % accuracy is obtained for Kannada, Telugu, Tamil and Malayalam numerals. M. Hanmandlu et.al proposed zone based feature extraction for handwritten Hindi numerals. The image is divided into 24 zones for feature extraction. Bottom left corner of the image is considered absolute reference, the distance vector for each pixel present in the zone is computed with respect to reference zone. Then normalized distance

vector is then computed by dividing the sum of distances vector of all black pixels in the zone with their total number and the process is repeated to obtain 24 feature sets.

Al-Taani Ahmad et.al, [6] presented structural method for recognizing on-line handwritten digits, input strokes are used for calculating and normalizing slope values of input coordinates. The change of direction is recorded using the successive slopes values. Finite Transition Network that contains the grammar of the digits is used to match primitive's string with corresponding digit to recognize the digit. The method is tested on sample of 3000 digits written by 100 different trained persons.

Chan et al. [7] presented a structural approach for recognizing on-line handwriting. Structural features are extracted from the input strokes. The presented approach on 62 character classes (digits, uppercase and lowercase letters) and each class has 150 entries. Experimental results showed that the recognition rates were 98.60% for digits, 98.49% for uppercase letters, 97.44% for lowercase letters, and 97.40% for the combined set.

There are still many issues in Urdu, Persian, Arabic and Old Arabic numerals recognition. The main issue is the lake of standard for handwritten Urdu script. In Urdu, numerals are written in both script Urdu old Arabic forms, mostly old Arabic is commonly used while Urdu is used only when it is written in standard format. The separation of numerals from the words still have big problem. A little research has been done for the separation of numerals from the Urdu, Farsi, and Arabic words. The problem of separation and recognition of old Arabic and Urdu numeral is resolved in this paper. This paper describes the similarities and dissimilarities between the two main used in Urdu numerals writing are Urdu (like Farsi) and old Arabic from the character recognition point of view.

II. PROPOSED SYSTEM

An online numeral recognition system is presented which is capable of recognizing the both Indian and Arabic scripts written through the digitizing tablet or light pen. The system is divided into four parts as shown in figure 3. A structural feature based approach is presented to identify on-line handwritten Urdu Script numerals which depend upon the directional features.

The input strokes are converted into image to process the strokes from offline domain for skew and slant correction and normalization. As handwritten strokes are not uniform and slant may change on some stroke on one line therefore instead of global slant correction that does not guarantee that all the strokes are corrected properly. Hence local slant correction [11] is used by divide the strokes into smaller size unit and perform slant correction on these smaller units. Similarly skew detection is performed through Hough transformation [11], strokes are divided into smaller unit i.e. two to three strokes instead of single stroke to get better skew correction, and not more than three strokes to perform locally skew correction.

The normalization process of the input strokes is essential because of the different writing styles and font which results in several variations in size and shapes of the strokes. Thus to attain proper recognition result input strokes should be normalized. Normalization is performed by defining a uniform grid in the offline domain.

Smoothing [8] is preformed to reduce the variation for better feature extraction. Baseline information is used for many purposes in handwriting recognition which represents orientation in a word and it is essential for many handwriting task i.e character recognition, personality identification, and writer identification [8]. The algorithm used in [8] failed to find the base line properly. Minimum enclosing rectangle is used to find the base line for more than one stroke as shown in figure 4.

The main purpose of feature extraction from the input strokes is the extraction of those distinct patterns that uniquely define the stroke and are most important for classification. The task of human expert is to select those features that allow effective and efficient recognition. In the proposed online handwriting numerals recognition system we focused on the structural features. Structural features include loops, cusp, endpoints, starts points etc. Structural features are the shape defining features and these are based on the instinctive aspects of writing. Due to similarity between numerals of structural features are used. Several shape defining features are extracted as shown in figure 5.

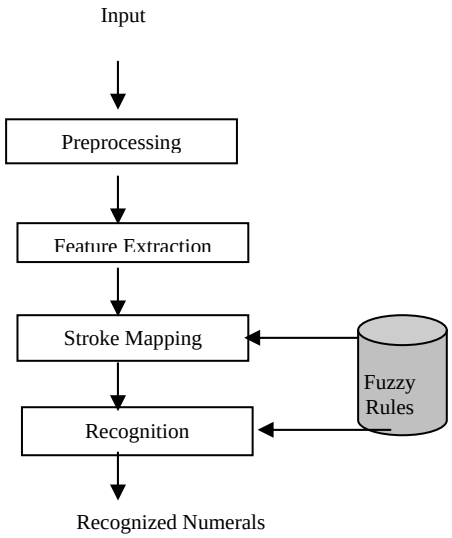

Figure 3: Basics structure for numeral recognition

Cusp is the sharp turning point exists in some numerals as shown in figure. The cusp is divided into three main categories up, down and right. Up cusp is extracted when movement of pen is from downward after upward, while down cusp is upward after downward. The right cup is extracted when sharp turning point exist with movement left before right.

The starting and ending directional features are extracted based on the chain code. Directional features

are divided into four categories are up, down, right, left, diagonal right downward, diagonal left downward as shown in figure.

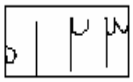

Figure 4: Minimum Enclosing rectangle to find the base line

Loop is an important feature to differentiate some similar numerals, loop is divided into two categories small (only for zero written in Urdu Numerals) and large for other numerals.

As for Urdu numerals, loop is very small for zero digit, small loop may occur between some numeral due to noisy input as shown in figure.

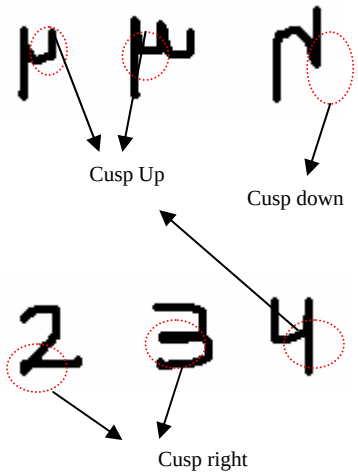

Figure 5: Cups categorization up, down and right

Feature purification process discards the incorrect features by using some defined rules. As small loop exist only in zero written in Urdu digits, thus small loop is removed from other feature matrix.

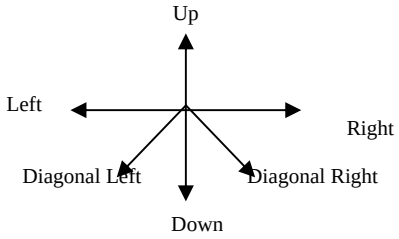

Figure 6: Directional features at starting and ending

A. Rule Based Approach.

Fuzzy provides a powerful tool for pattern recognition of irregular patterns. Fuzzy linguistic are the formal representation of recognition system made through fuzzy if/then rules. The core of the recognition process is the if/then fuzzy rules. And the features are the input to the fuzzy rules. The features are extracted with respected to timing as they occur. The inputs to the rules are features and timing information of each feature. These features are encoded using the fuzzy if then rules shown below.

Following rules are defined for the recognition of numerals five in Urdu script.

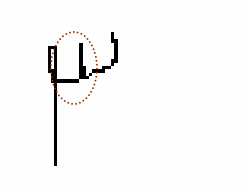

Figure 7: Small loop issue

If starting and ending points are same and only one loop exist.
If up cusp exit
Then five in Urdu
If single stroke to recognize
The zero in Old Arabic
if size-defined size of small <
size – defined size of large
then zero in Urdu script
if size-defined size of small >
size – defined size of large
then five in Urdu script

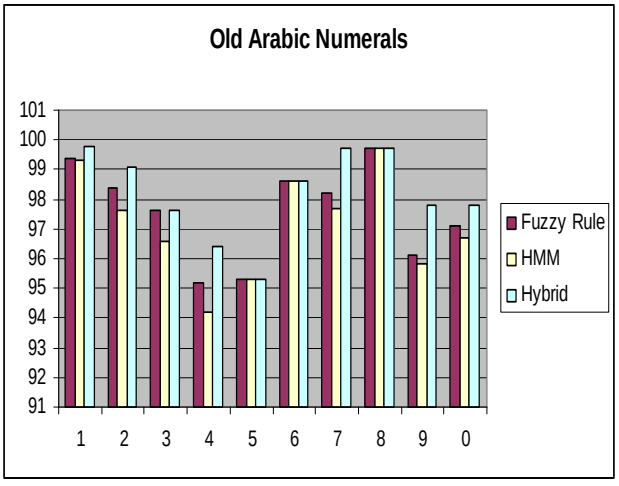

Figure 8: Recognition rate of old Arabic numerals

The description of the above rules is if the starting and ending is same then loop exist, the starting and ending of the numerals are zero both in Urdu and old Arabic, five in Urdu, and eight in old Arabic. If it consist of two loops then the stroke is eight, else if the cusp exist between ending and starting point then clearly it five in Urdu

script. If the written stroke is only one to recognize then it is considered as Old Arabic zero. If size of current stroke is less than the defined size of zero in Urdu then it is close to zero written in Urdu script and vice versa.

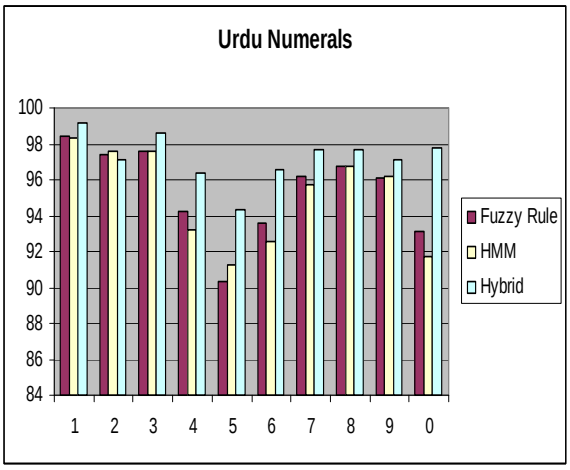

Figure 9: Recognition rate of Urdu numerals using fuzzy logics

B. HMM Based Approach

Separate HMMs is built for each numeral which have 4 states and 23 observation symbols. Features matrix is the input to the HMM as a observation symbols. For observation symbols 12 structural features are extracted. These structural features are eight small length directional features i.e in numerals 3 and eight long length directional features i.e in numeral 1 and 7 obtained from the chain code, loops, intersection, cusp upward and cusp downward, curve right, left, upward.

C. Hybrid Approach

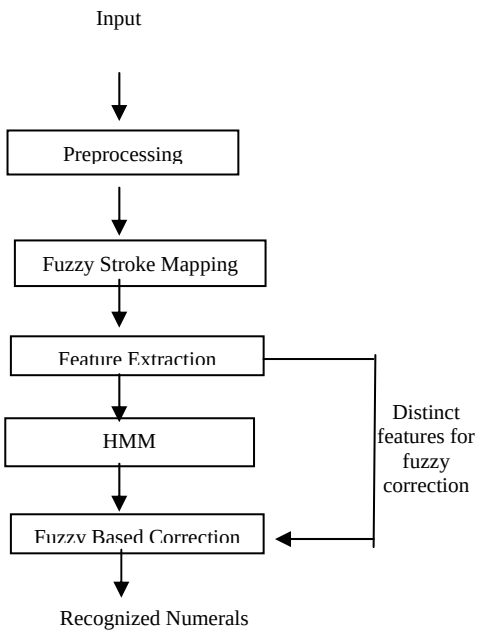

Figure 10. Hybrid Approach

For some numeral HMM failed where there is the small difference between Indian numerals like 1 and 7, 6 and 8, 4 and 9, similarly for old Arabic numerals. This confliction is resolved using the hybrid approach through fuzzy logics and HMM. A post processing step is applied that future decides upon the structural features and posterior probability obtained from the HMM. Its takes the all the identified ligatures with probability with in the threshold, and then identified the stoke with the help of unique features at starting and ending of the character. If there are no unique features that are helpful in defining the shapes of the output probability form the HMM then the decision is based on the probability form the HMM.

For Ligature L,
Select HMM with maximum Probability L_{HMM}
For all other L_i *where i=1,2,…. 22*
Calculate the probability difference $L_{HMM} - L_i < \alpha$
For $L_s = L_{HMM} - L_i < \alpha$
Apply Rules for Post Processing
Hybrid Approach Rules below
Otherwise L_{HMM} *is the recognized ligature*

Hybrid Approach Rules
For all ligatures L_s *Repeat the following steps*
If L_s *=1 and 7*
If there is small left directional features at starting then $L_s = 7.$,
Else if there is small left to right movement at starting and ending is downward with little diagonal then $L_s = 7$
Else consider ligature $L_s = 1.$
Similarly Rule for All other Strokes L_i

III. Conclusion

This paper presented the similarities and dissimilarities between these two scripts Urdu and old Arabic from the character recognition point of view. Rule based technique, HMM and Hybrid approach is presented to recognize the online digit recognition written in both Urdu and Old Arabic forms from both online and offline domain. As very little research has been done for the separation of numerals from the Urdu, Farsi, and Arabic words. The problem of separation of old Arabic and Urdu numeral is resolved in this paper. The digits are written either in Urdu or Arabic but not both at the same time. The system provides accuracy of 97.4% using fuzzy rule and 96.2 using HMM and 97.8 using Hybrid approach on 900 samples by taking the input from 30 trained users. There are still many problem exist in Urdu script due to complexities in the script. The proposed technique work only for numerals input. The separation of numerals from the Urdu text still have big problem due to shape similarity.

V. REFERENCE

[1] J. Sadri, et.al , “Application of Support Vector Machines for recognition of handwritten Arabic/Persian digits,” Proceeding of the Second Conference on Machine Vision and Image Processing & Applications (MVIP), Vol. 1, Feb. 2003, Iran, pp. 300-307.

[2] Harifi et.al, “A New Pattern for Handwritten Persian/Arabic Digit Recognition”, International Journal of Information Technology, Vol 1, Number 4, pp 174-177

[3] Plern Kortungsap1 et.al, “On-line Handwriting Recognition System for the Thai, English, Numeral, and Symbol Characters

[4] S.V. Rajashekararadhya et.al “Efficient Zone Feature Extraction Algorithm for Handwritten Numerals Recognition of Four South Indian Scripts”, Journal of Theoretical and Applied Information Technology 2008

[5] M. Hanmandlu et.al , “Input fuzzy for the recognition of handwritten Hindi numeral:”a, *International Conference on Informational Technology* 2007.

[6] Al-Taani Ahmad et.al, “Recognition of On-line Handwritten Arabic Digits Using Structural Features and Transition Network” Informatica 2008.

[7] Kam-Fai Chan and Dit-Yan Yeung. Recognizing on-line handwritten alphanumeric characters through flexible structural matching. *Pattern Recognition,* Vol 32, pp. 1099 - 1114, 1999.

[8] M.I.Razzak, Muhammad Sher, S.A.Hussain, Z.S.Khan, “Combining online and offline preprocessing for online Urdu character recognition” IMECS 09.

[9] M. Pechwitz, V. M¨argner, “ Baseline Estimation For Arabic Handwritten Words”, IWFHR’02.

[10] Javad sadri et.al “State of the art in Farsi script recognition” Signal Processing and its application, 2007.

[11] Faouzi Bouchiareb, Mouldi Bedda, Salim Ouchetai "New Preprocessing Methods for Handwritten Arabic Word" Asian Journal of Information Technology.