Combining Offline and Online Preprocessing for Online Urdu Character Recognition

Muhammad Imran Razzak, Syed Afaq Hussain, Muhammad Sher, Zeeshan Shafi Khan

***Abstract*—Urdu online handwriting recognition is a very challenging task due to its cursive nature. Pre-processing of the raw input strokes is crucial part for the success of character recognition system. The findings from online data are not enough for recognition of Urdu due to the complexities of Urdu script. This paper describes the preprocessing steps for online character recognition. A novel technique is presented for preprocessing of Urdu online text in which both online and offline domain are used to remove the variations and to increase the efficiency of the recognition system for online input. The proposed technique is also the necessary step towards character recognition, person identification, personality determination where input data is processed from all perspectives.**

***Index Terms*—Online Preprocessing, Online Urdu Character Recognition, Preprocessing, Urdu Character Recognition**

I. Introduction

Automatic handwritten recognition has been classified into two categories based on the input data: online and offline. Offline handwritten recognition does not require immediate interaction with the user while online handwritten recognition has completely interaction with the user. The root of online handwriting recognition is real time data collection by way of a digital sampling method. The most common input devices are digitizing tablets or digital pen, where the written data is digitized and translated into a series of coordinates. . For on-line recognition, a digitizer samples the handwriting to time-sequenced pixels as it is being written. Hence, the on-line handwriting signal contains additional time information, which is not presented in the off-line signal. Pen-based input gives a lot of advantages. First of all, it helps users, such as computer novices and old people, to conveniently use a computer. It also makes a small size portable computer (PDA, handheld PC, palm PC, etc.) possible because there is no need for keyboard or keypad. The data is input through some Digital Pen, Writing Board and Styles.

Chinese and English are languages, which have tremendously attracted interests of character recognition researchers both for online and offline while in contrast, research efforts in the field of character recognition for Urdu, Persian and Arabic scripts face major problems . These problems are due to complexities of this script like cursiveness, multiple shapes of one character at different positions in a ligature, overlapping and connectivity of characters on the baseline.

Urdu Script contains 58 characters and more than 10 commonly used fonts i.e. Nastaliq, Nasakh, Noori Nasakh, Noori Nastaliq, Koofi, etc. Nastaleeq and Nasakh are the two most popular fonts. Nastaleeq is a special calligraphic way of writing and is mostly used especially for handwriting. It does not have a baseline rather the text is centre justified and it is very difficult to recognize this style of writing because of its complexity and Nasakh is another writing style of Urdu which follows one baseline. Thus this way of writing is simpler than Nastaleeq so it is easy to recognize this style because of its simplicity.

Pre-processing phase is basic and important step towards the success of character recognition system so variation can be removed and data will be ready for feature extraction. As Urdu script base languages are very complex as compared to the other languages thus it is very difficult to obtain a reasonable accuracy only trusting on the online data.

The basic aim of this study is to propose an offline processing in online Urdu character recognition that runs parallel with online processing. This proposed method applies both on-line and offline preprocessing to the input online data so that finding requires to recognition engine can be increased. Section II introduces the literature work, in section III proposed solution is discussed, finally finding and discussion are discussed in section IV.

II. Literature Work

Within the context of online handwritten character recognition, studies dealing with Arabic characters are scarce [1]. The main task in preprocessing the handwritten data is to decrease the variation that causes to reduces the recognition rate and also increase the complexities. Pre-processing of the input raw stroke is crucial for the success of efficient character recognition systems.

The main objective of the preprocessing steps is to normalize words and remove variations that would otherwise complicate recognition and reduce the recognition rate. S.A. Husain et.al.[2] perform smoothing and de-hooking before feature extraction as a preprocessing phase. As the input text contain irregularity due to the natural way of writing, they perform 2 to 3 pixel smoothing and de-hooking on the Urdu input text. S. Malik and S.A. Khan [3] performed repetition removal and filtering step in preprocessing phase. Repetition Removal: One of the more important things to remember

Manuscript received December 20, 2008. This work was supported in part by the Department of Computer Science, International Islamic University, Islamabad..

Muhammad Imran Razzak is doing PhD from International Islamic University, Islamabad, Pakistan (e-mail: imran.mian@yahoo.com).

Syed Afaq Hussain is with the Department of Computer Science, Air University, Islamabad, Pakistan (email: drafaq@gmail.com).

Muhammad Sher is with the Department of Computer Science, International Islamic University, Islamabad, Pakistan. (e-mail: m.sher@iiu.edu.pk).

Zeeshan Shafi Khan is doing PhD from International Islamic University, Islamabad, Pakistan (e-mail: zeeshanshafikhan@yahoo.com).

about inking is that the digitizer that measures pen movement is very accurate. The grids used in these digitizers have 1,000 lines per inch. As slight movement of the pen tip is detected, even when the tip is not quite touching the digitizer but is hovering over it. So this will result in repetition of points and their presence can slow the processing speed of recognizer. Filtering is used for the removal of the noise i.e jitter and smoothing of the input sequence while M. Hussain [4] calculates the only displacement from 4,8,16 displacement between two points. K. Al-Ghoneim[5] performed translation, scaling, connected line generation, and smoothing. They compute the image's center of gravity, and translate the image such that its origin is the center of gravity. Scale the image so that the maximum radius for the character pixels equal to half the grid size. The radius of a pixel is defined as the length of the straight line connecting the pixel to the origin and connected line generation: they used the Bresenham's line algorithms to fill the missing points that was left due to fast pen movement and finally, smooth the input curve by inspecting each subsequence of pixels, and replacing it by a shorter version. Fadi Biadsy[6] worked on geometrical processing phase to minimize handwriting variations. They used low pass filter algorithm to reduce noise and to remove the imperfections caused by acquisition device. To eliminate the redundant points that are irrelevant for classification Douglas and Peucker's algorithm [7] was used. The document is broken into text lines and to words. The handwriting text line extraction techniques for On-line, depending on the y-axis histogram projection and character geometry (width, height, etc.) does not function well on Arabic handwriting due to its characteristics. Thus, they worked on new technique that is more suitable to the Arabic language nature [8]. A stroke is defined as all data-point samples drawn/written between a certain pen-down (Start of writing action) action and the following pen-up (Lifting the pen up after writing) action. Thus, a stroke may represent one or more character, or even a part of character, or sometimes a dot. By examining the states of successively written Arabic strokes (either main-type strokes or complementary-type like dots for example) that they are related spatially to each other by one of the following relations. This technique is basically for word formation. The system presented [9] incorporates normalization for each of the following factors: Slope, Stroke width, and height of letter. The normalization task reduces each letter to uniform height on horizontal base line and slant and slope is corrected.

The study presented [2-4, 6-9] only performed preprocessing only on the online side are smoothing, and de-hooking [2] and connected line generation [5] while somaya [9] worked on normalization of strokes. The existing techniques are on online information while they did not focus the offline processing for online data, which are base line finding, combining strokes and skews correction ect. Due to the complexity of Urdu text, it is better to utilize the offline processing along with online processing for online input text.

III. Proposed preprocessing

Pre-processing of the raw input strokes is crucial for the success of efficient character recognition systems. It takes a raw input strokes that contain *x,y* coordinates, timing, force and enhances it by reducing noise and distortion, variations and hence simplifies the strokes for next processing. The aim of pre-processing step is to process the input strokes so that variation can be removed and forward it to the next phase for accurate feature extraction and for efficient recognition rate.

In this paper we performed different preprocessing steps on the input strokes from both online and offline perspectives to utilize as much as information possible for recognition. This involves stroke segmentation, interpolation, smoothing, de-hooking and combining strokes, base line. The Online preprocessing steps are stroke acquisition, stroke segmentation, smoothing, interpolation and de-hooking. While the offline preprocessing steps are stroke combining and baseline finding. Offline processing is introduced parallel with the online because it is requirement to achieve better accuracy for Urdu script.

A: Online Preprocessing:

The input strokes are obtained through the tablet pc (100 DPS) or digital pen. These strokes consist of *x,y* coordinates, timing, force which will be used for preprocessing. As every writes the character with different force, it is mostly used for person identification, personality determination ect but it is not suitable for online Urdu character recognition. The following are the online preprocessing step applied on input stroke.

Stroke Segmentation: As the basic rule is that any Urdu character has one main stroke and zero or more secondary i.e. diacritical marks ect. strokes without and with diacritical marks as shown in figure 1. Each word of Urdu can be broken into ligatures. . Each ligature comprises of one or more than one strokes (primary stroke and secondary strokes). The combination of these ligatures forms different words. Each ligature consist of one base strokes and zero, one or more than one secondary strokes as shown in figure 1. Segmentation process segments the primary stroke and secondary strokes. The proposed segmentation is based on the threshold value and position with respect to the previous base character. Basically the position is important for some secondary strokes where secondary storks are of greater size like in secondary stroke tawn ط as shown in figure 1-(b) because it is of same size when compared with the size of some primary stroke ط and ر . Thus only utilizing stroke size it is very difficult to segment these secondary strokes. So some additional information, the position of the secondary stroke with respect to primary stroke is used to segment the primary and secondary strokes. The basic rule used is that first written character is compared with the threshold, if it is less than threshold value θ_1 then it is considered as secondary stroke and if it falls with the threshed value θ2 then it position is calculated with respect to the previous strokes to check weather it falls with in criteria i.e up, down or with in θ_3 (where θ_3 heck weather the character is within range of previous stroke like for لا.

ب ا

Figure 1 (a): Secondary Strokes and Primary Strokes

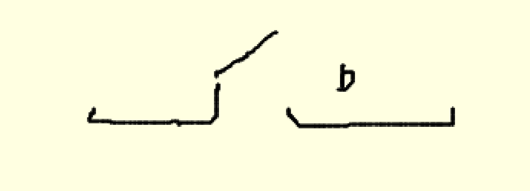

Figure 1(b) : Secondary strokes with greater size

Interpolation: Due to the limiting processing power of pen and low camera frame rate, it skips some points as shown in figure 2; this depends upon the writing speed, to compute the missing point's interpolation is performed for correct classification. More over for better result the writer should write with normal speed. Some important feature missed when high speed text was written by user for training purpose. We have used Bresenham's line algorithms for interpolation that estimate the missing intermediate points.

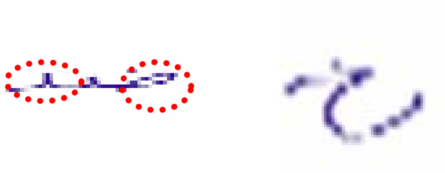

Figure 2: Missing points due to fast speed of writing.

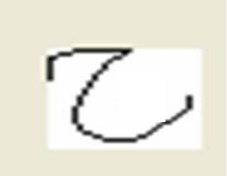

Figure 3: Interpolation of Missing points.

Smoothing: As the input strokes contain zigzag path due to naturally hand shivering during writing. Although this is the slow variation, but this may effect on the recognition rate. Smoothing is one of simplest approaches for data filtering. It consists of substituting the coordinates of the original point by using the neighboring points. In this work smoothing was done on the chain code of the stroke because mostly chain code are used to extract features instead of *x,y* coordinates. Three pixel smoothing was done as per the variations in the chain codes of the stroke.

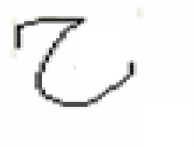

Figure 4: Smoothing on stroke shown in figure 3.

De-hooking: Hooks are very common artifacts found at the beginning and end of the strokes. As the pen is very sensitive thus they are generated during fast writing or writing by inexperienced person, when pen-down and pen-up events are generated. These often create problems in the detection of the original ligature. Therefore, it is very important to remove them. These usually occur at the beginning and the end of the stroke. The hooks occurring at the beginning and the end of the stroke are removed with the help of the generated chain codes. If variation in the chain code (last 6 chain code in length) at the beginning or end is less then the specified threshold, then that part considered a hook and is removed by either discarded it or replacing the respective co-ordinates with the neighboring ones. Hooks are generated by users either he is experienced or inexperienced. As there is a very small lines that is added at start and end but some problems exist in removing these hooks, it may possible that small up of Jeem is removed as instead of hooks which is the most important part in the detection of jeem as shown in figure 6.

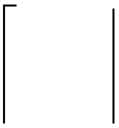

Figure 5: Hooking and De-hooking on Alf

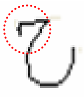

Figure 6: Hook in Jeem

So to avoid de-hooking in jeem, de-hooking at beginning is not performed on those ligatures which are written from left to right for some length like ج. The isolated jeem is written from left to right and the ligature جر is also written left to right at beginning while the remaining ligature is written from right to left.

B: Offline Preprocessing:

In this phase input stroke are transformed into image to perform offline preprocessing steps. Only online finding are not enough for Urdu online character recognition due to its characteristics. It is more suitable to involve offline information along with online information to increase the recognition rate.

Ligature Combination: It is difficult to write some ligatures i.e بصر, لا ect. without lifting the pen as shown in figure 7, while online character recognition does not permit to lift pen during writing. Thus to overcome this issue, a new algorithm is proposed that treat the two consecutive ligature as one ligature if they overlap each other and starting point of second ligature is close to the ending of the first stroke. This value threshold is considered twice in vertical than in horizontal. If the two ligature form single ligature using offline approach, then these two ligatures are combined and considered as a single ligature as shown in figure 8.

Base Line: Baseline information has been used for different purposes in handwriting recognition. The baseline represents a first orientation in a word. This line represents a orientation in a word and is necessary for many handwriting task i.e personality identification, writer identification. The algorithm [10] is used to detect baseline.

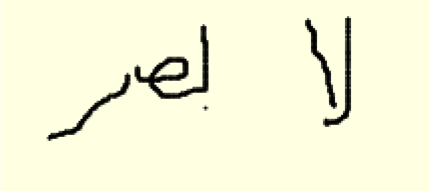

Figure 7: Base storks that are difficult to write in single stroke

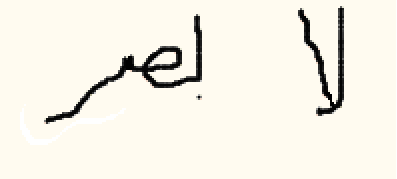

Figure 8: Resolved problem shown in figure 7.

IV. Conclusion and discussion

As pre-processing of the raw input strokes is crucial phase for the success of efficient character recognition systems. The findings from online data are not enough for recognition of Urdu due to its complexities. This paper describes the preprocessing steps for online character recognition. Due to the complexity of Urdu script, a novel technique is proposed for preprocessing in which both online and offline preprocessing are used to remove the variations and to increase the efficiency of the recognition system for online input. By using the joint processing for online the efficiency can be increased. The proposed technique is the necessary step person identification, personality determination where input data is processed from all perspectives.

V. Future work

This implementation is initial step. Therefore, there is a lot of scope for future enhancement, i.e. implementation of other preprocessing techniques i.e. text rotation, normalization ect. feature extraction techniques. This is also the first step towards the person identification and personality determination based on Urdu text where the size, orientation and position of characters relative to each other are very important.

.

References

[1] Inam Shamsher et.al, OCR For Printed Urdu Script Using Feed Forward Neural Network, Proceedings of World Academy of Science, Engineering and Technology. Vol 23, Aug 2007 ISSN1307-6884

[2]. S.A.Hussain , Anwar F., Asma. "*Online Urdu Character Recognition System*." MVA2007 IAPR Conference on Machine Vision Applications.

[3]. Malik, S. Khan, S.A., "Urdu Online Handwriting Recognition", Emerging Technologies, 2005. Proceedings of the IEEE Symposium on Volume, Issue, 17-18 Sept. 2005 Page(s): 27 – 31,

[4].M. Hussain et. al. " Urdu Character Recognition Using Spatial Temporal Neural Network", IEEE 2006.

[5] K. Al-Ghoneim et.al. "Sub Stroke Approach to HMM-based On-line Kanji Handwriting Recognition**"** , IEEE 2001.

[6] Fadi Biads et.al, *"Online Arabic Handwriting Recognition Using Hidden Markov Models"IFFHR10 2006..*

[7] D. Douglas and T. Peucker. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *The Canadian Cartographer*, 10(2):112–122, 1973.

[8] Randa I. Elanwar, Simultaneous Segmentation and Recognition of Arabic Characters in an Unconstrained On-Line Cursive Handwritten Document, International Journal of Computer and Information Science and Engineering Volume Number 4.

[9]. Somaya Alama' adeed et.al. Recognition of Offline Handwritten Arabic Word Using Hidden Markov Model Approach , IEEE 2008

[10]. M. Pechwitz, V. M¨argner, " Baseline Estimation For Arabic Handwritten Words", IWFHR'02.