MVA2007 IAPR Conference on Machine Vision Applications, May 16-18, 2007, Tokyo, JAPAN

Online Urdu Character Recognition System

S. A. Husain, Asma Sajjad, Fareeha Anwar
drafaq@iiu.edu.pk
Dept of Computer Science, Faculty of Applied Science, International Islamic university, Islamabad.

Abstract

Handheld devices generally provide the facility of text input through keys that are an inconvenient and slow way of input. Digitizing tablets and light pens, on the other hand provide a natural and convenient way of input. There are many online character recognizers for languages based on Roman and Chinese characters but there is no such commercial product for Urdu/Arabic text input. We present the design of an online Urdu handwriting recognition system that can recognize about 850 single character, 2 character and 3 character ligatures, enabling input of about 18000 common words from the Urdu Dictionary.

Keywords: Online Character Recognition, Urdu handwriting recognition, Ligature based identification, digitizing tablets, handwritten characters, feature extraction, Back-Propagation Neural Network.

1.INTRODUCTION: The user interface is a means by which people interact with a particular machine or other system. Efforts are being made on both the software and hardware side to make this human computer interaction more and more friendly.The development of pen interfaces is a key element in providing an efficient and natural way of input to the computer for e.g. PDAs usually have a graphical user interface in which a pen can be used for pointing and selection functions, drawing, and text entry. Urdu is the national language of Pakistan. The hand held devices have also successfully emerged in Pakistan but the software they provide for user input are mostly in English. Where as the common man in Pakistan can not communicate in English easily. In order to reduce this difference between the common man and the new technology Urdu input software were required. Our research is a step in order to bridge this gap. Urdu handwriting recognition is a complex process due to the complexity of the Urdu script as described below:-**Cursive:** Urdu text is cursive in nature [10]. Adding to the complexity is the writing style in which the characters forming words are connected to each other. **Ligatures:** Several characters of Urdu are combined vertically to form a ligature [10]. **Spaces:** Spaces in Urdu may occur between ligatures and between words. The spaces between ligatures and words vary. This also makes the recognition difficult. **Overlapping:** Recognition of individual characters with in a ligature becomes quite difficult as the characters in Urdu overlap vertically and do not touch each other. **Diacritics:** Diacritics are very important in Urdu language. These include diacritics such as: *Dots, Tay, Hamza, Diagonal & Mad*, etc. [16]. **Context sensitivity:** Every character in Urdu can have up to 4 different shapes (in Nasakh Font) depending on its position with in a ligature i.e. whether the character is isolated, in the beginning, at the end or connected from both sides in a word [10]. **Strokes:** The basic rule is that any Urdu character has one main stroke and zero or more secondary stroke **Direction of writing:** Unlike English, Urdu is written from right to left. [10]. **Presence of a Base line:** Urdu has a base line. The base line is a horizontal line which runs through the text, cutting all the words at some point.

It is the complexity of Urdu script which poses great challenges to the new field of online Urdu handwriting recognition.

2.PREVIOUS WORK: There are many offline OCR systems available for printed Arabic/Urdu documents. However, there are rare online (dynamic) OCR systems for Urdu language. This may be due to the complexities involved in the online character recognition with the added difficulties of Urdu handwriting. There are basically two techniques for recognizing words. The segmentation based which involves the division of a word in to its individual characters. Other is the segmentation free or ligature based recognition, in which the word is recognized as a whole without trying to segment it in to characters.Malik and Khan [18] have recognized only individual characters and Urdu numerals but ligatures have not been addressed. Using the individual

characters, 200, two character words were recognized. For example, ب ا, ر د, etc. They have used tree based dictionary search for the classification of characters. The recognition rate for isolated characters and numerals is 93% and 78% for two character words. Another work reported in this field is a research project completed at NUCES [19]. They have recognized words using the segmentation based approach. Characters were classified into 60 classes. STNN was used for recognition. The segmentation based approach is only valid for the Nasakh Font, and even the large number of class works for a small dataset. The earlier works are too primitive as encompassing only isolated characters. Our work is a considerable improvement with ligature based approach applying to a larger dataset and is independent of font (script).

3. PROPOSED METHODOLOGY: Due to the cursive nature of the Urdu handwriting, recognition was difficult. For the popular Urdu Nastaliq writing style, we have used the segmentation free approach. Here, each input stroke represents a ligature which is not broken in to characters as many of the recognition errors occur due to segmentation. The segmentation free system extracts a feature vector for each ligature which is then passed, to the BPNN for its classification. Using the stroke(x, y) co-ordinates and the chain codes, unique features for every stroke are detected and a feature vector is extracted. This feature vector is then fed in to the BPNN for the classification of every stroke in to its respective class. In Urdu, a number of secondary strokes are utilized which do not mean anything in isolation but are associated with a ligature to give it a meaning, just like a dot in English. The following secondary strokes are recognized:

Figure 1: Secondary Strokes

Namely, (left-right) small tuan, hay, long diagonal stroke, Madaa, Hamzaa, and the single dot. These special ligatures are associated with the base ligature. After this, the ligature is checked for its validity. Valid ligatures form words. After word formation, word validity is checked by using a word dictionary. Finally, the valid words are written in a text file. The OLUCR recognizes 38 one character ligatures, 709 two character ligatures e.g. حب, شا etc and approximately 50 most commonly used three character ligatures e.g. تھا , بھی , لتا , ٹیپ etc. The **constraints** for the system are that the base stroke should be written before the secondary stroke. For the 2 character ligature expecting a secondary stroke, the secondary stroke for the first character should be written first and then for the second character.

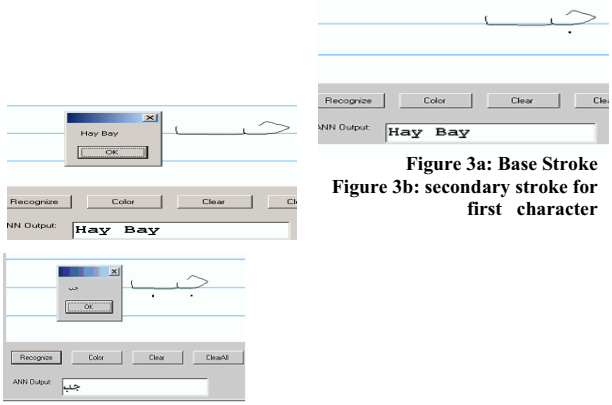

Figure 3a: Base Stroke
Figure 3b: secondary stroke for first character

Figure 3c: secondary stroke for second character

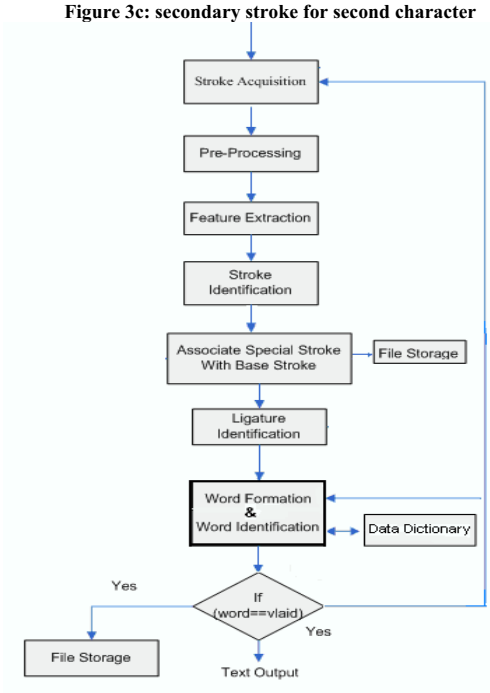

Figure 2: Block Diagram of Proposed System

3.1 System Modules : Keeping the challenges of Urdu online character recognition in mind, our system consists of the following modules which are also the building blocks of many online character recognizers.

3.2.1 Acquisition The processing was done on strokes obtained using digitizing tablet, the Intuos Wacom board. The data is collected in real time. The standard data is a stream of {x, y} coordinates. The data resolution is around 100 DPS (dots per second).

3.2.2 Preprocessing The data obtained often contains irregularity i.e., the hooks and erratic handwriting generated by inexperienced users. Hooks occur due to the inaccuracies during pen up/pen down while placing the stylus on, or lifting it off the tablet. These were removed using 2 to 3 pixel smoothing.

Figure 4: written by inexperienced writer and containing hooks in the beginning and end.

3.2.3 Feature Extraction This stage extracts distinguishing features for base ligatures and secondary strokes. For the base ligatures a feature vector consisting of 20 features was prepared. The features extracted were syntactical i.e. they identified various shape forms present in the Urdu ligatures i.e., loops in the beginning or end, intersections, direction/writing style of any ligature. These also included features that are selected on the presence of certain alphabets of Urdu language. For example there is feature which is selected on the presence of in any ligature for e.g. . These features were very helpful in uniquely distinguishing the ligatures.

3.2.3.1 Features for Base Stroke

Start Vertical: This feature was selected when the ligature was a straight vertical in the beginning e.g. . **End Vertical** This was selected when the ligature was a straight vertical in the end e.g. . **Horizontal R2L** If while writing the ligature the pen movement is from right to left horizontally then a bit is set in feature vector e.g. . **Horizontal L2R** While writing the ligature, if the pen movement is from Left to right horizontally then this feature is set in the feature vector e.g. . **Hedge** In Urdu characters like, a semi circle sort of shape is present which we call curve. For such characters we have selected a feature called the hedge. **CurveR2L** The direction of writing of these curves varies from right to left and also from left to right. Therefore, Curve R2L has been set for characters whose writing direction is right to left e.g. . **Curve L2R** If the curve direction of the character from left to right e.g. , . **Loop Flag** Loops are very common features of Urdu handwriting. They are present in characters i.e., . If the recognition engine finds a loop it selects this feature. **Cusp** A cusp is a sharp turning point in a stroke. This feature is selected for the ligature which contains the cusps such as those present in and as shown in the figure below.

Figure 5: Cusp in character

Intersection When ever an intersection is encountered in a stroke this feature is selected e.g. in etc. **Ray** This feature is selected for the character ray of Urdu alphabet. If any ligature is a combination of ray then this feature is also selected for that particular ligature e.g. etc.

End Up Vertical This feature is selected for ligatures having a vertical end in the upward direction e.g. etc. **Loop Up** The writing direction of the loop in ligatures started with is from down to up as shown in figure below.

Figure 6: Writing direction of loop of م

Seen Bit This feature was selected if character seen is detected in any ligature e.g. **Aien Bit** This feature was selected if character is detected in any ligature e.g. **Hay Bit** This feature was selected if **Hay** is detected in any ligature e.g. **Dal Bit** If the recognition engine detects a د in the ligature written it selects this feature **Double Loop** This feature is selected for ligatures which have two loops e.g. ligatures like

Figure 7: double loop ligatures

Tuan Bit This feature is selected on the presence of ط in any ligature e.g. **.Gol Hay** This feature selected when gol hay is at end e.g. etc. The shape of gol hay ligatures are

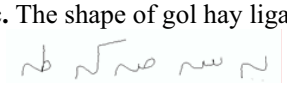

Figure 8: Gol Hay Ligatures

3.2.3.2 Features for Secondary Stroke:

Dot If there is a dot with in the boundaries of base stroke this feature is selected. **Madaa** If there is a madaa with in the boundaries of base stroke this feature is selected. **Diagonal** This secondary stroke feature is selected for the diagonal stroke occurring in ک and گ over a base ligature.

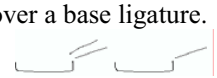

Figure 9: The diagonal stroke over ک and گ

Hay This feature is selected if the secondary stroke called the hay is encountered. e.g. in .

Figure 10: The hay stroke

Hamzaa It is a stroke which is present over the base strokes. If the secondary stroke is Hamzaa then this feature is selected. **Chooti Tuan** It is present over the base ligatures. If the loop follows a vertical line then the stroke is ظ.

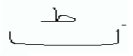

Figure 11: The ط over the base stroke kashti.

3.3 Stroke Identification: This stage involves the identification of the base stroke and then that of the secondary stroke. This identification has been done using BPNN. The network for the base ligature recognition consists of 20 inputs, 153output and 69 hidden nodes. The learning rate for this has been set to 0.6F. The number of hidden nodes and the learning rate were set through testing various values. Secondary strokes recognition consists of 5 inputs, 5 outputs and 10 hidden nodes. The learning rate for this has been set to 0.5F.

3.4 Ligature Validation: In this, the ligature written is checked for its validity. If ligature is valid then the next ligature is considered. Otherwise, ligature is removed to be rewritten.

3.5 Ligature Combination/ Word Formation: In this stage the difference between ligatures is considered. If the difference between the adjacent ligatures is less then the threshold which is necessary for word formation then these ligatures form one word otherwise the new ligature is taken up as another word. In this way a series of words can be written. For example آ is one ligature and تا is another ligature. When we combine these two, the word آتا is formed.

3.6 Valid Word Identification: This stage makes use of the dictionary to identify that the word written is valid or not. If the dictionary does not contain the word then the word is discarded. The dictionary comparison is based on the words Unicode comparison. Invalid word is also not written in file.

3.7 Output: The output is the Urdu Text in the interface's text area and in a word file. The Unicode of every ligature verified is stored and once the word formed with the ligature is identified as valid it is written to a text file. The writing is done by sending Unicode to the file through the program.

آپ کام کرتے ہیں.

Figure 12: The text in file.

4. RESULTS: The system was trained using a training set of 240 carefully selected ligatures. With the combination of 6 diacritics we have successfully recognized more than 864 ligatures. These ligatures form approx 50000 words. The Recognition rate of base ligatures was 93% and of the secondary strokes was 98%. Test results of some of the difficult ligatures and diacritics (Aerab) are given in the tables below.

No	Ligature	Recognition Rate
1	٠	100 %
2	~	85 %
3	/	100 %
4	ِ	95 %
5	ـّـ	95 %
6	ء	95 %

Table 1: Recognition Rate of diacritics w.r.t. samples

No	Ligature	Total Samples	Recognition Rate w.r.t. samples in %	No	Ligature	Total Samples	Recognition Rate w.r.t. samples in %
1.	ب class	18	89.3	11.	حذ	16	95.5
2.	س class	18	90.6	12.	فص	16	95.5
3.	ص class	18	87.3	13.	فق	16	95.5
4.	ط class	18	93	14.	مق	16	95.5
5.	با	18	96.8	15.	مج	16	85.7
6.	طا	18	86.5	16.	عبے	16	94.5
7.	طب	18	87.3	17.	طو	16	94.5
8.	قب	18	87.3	18.	سبہ	16	91
9.	صح	16	83.5	19.	ذہو	16	87.5
10.	حم	16	95.5	20.	پش	16	96.5

Table 2: Recognition Rate of base ligatures w.r.t. samples

5 Concluding Remarks This paper, presents a method for recognition of online Cursive Urdu hand written Nastaliq Script. The system is currently trained for 250 ligatures. Our approach minimizes the errors by using segmentation free approach. By using multiple features, we have improved number of ligatures that can be identified. We have successfully recognized 250 base ligatures and 6 secondary strokes. These when combined form 864 ligatures which can recognize 50000 words of our Urdu dictionary successfully.

6 Future enhancements: This implementation was an initial step. Therefore, there is a lot of scope for future enhancement, i.e. implementation of other pre-processing techniques i.e. RTS techniques. Then enhancement in the number of ligatures which is a continuous area of research. i.e. the recognition of 4 characters ligatures and so on. The recognition of additional secondary strokes such as the shad, zeer, zabar and paish. Also, recognition of Urdu numerals.

REFERENCES

1. http://www.ethnologue.com/show_language.asp?code=urd.
2. http://www.omniglot.com/writing/urdu.htm.
3. http://www2.psy.uq.edu.au/~brainwav/Manual/BackProp.html
4. http://std.dkuug.dk/JTC1/SC2/WG2/docs/n2413-3.pdf.
5. Mohammad S. Khorsheed, William F. Clocksin, "Structural features of cursive Arabic script", proc of 10th British Vision, Conference, University of Nottingham, UK, September-1999

6. M Khorsheed,"OffLine Arabic Character Recognition A Review".

7. M S Khorsheed, "Automatic recognition of words in Arabic manuscripts", PhD Dissertation, Churchill College, University of Cambridge, June 2000.

8. H. Bunke, P. Wang, "Handbook of character recognition and document image analysis", World Scientific, 2000.

9. S A Husain and S H Amin, "A Multi-tier Holistic approach for Urdu Nastaliq Recognition", INMIC 2002.

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

11.Sutat Sae-Tang Ithipan Methaste. "Thai Online Handwritten Character Recognition Using Windowing BPNN", Information Research and Development Division, National Science and Technology Development Agency, Rachathewi, Bangkok, Thailand.

12.VKazushi Ishigaki VHiroshi Tanaka VNaomi Iwayama. "Interactive Character Recognition technology for Pen-based Computers".