Segmentation of Nastaliq Script for OCR

Sohail A. Sattar, Shamsul Haque, Mahmood K. Pathan

Department of Computer Science and Information Technology
NED Universiy of Enginnering and Technology, Karachi, Pakistan
Tel : 92-21-9261261, Fax : 92-21-9261255
E-mail :{ sattar, pvc ,deanish}@neduet,edu.pk

ABSTRACT

In this paper we have presented a novel segmentation technique for the implementation of an OCR (Optical Character Recognition) for printed Nastalique text, a calligraphic style of Urdu which uses the Arabic script for its writing. OCR for many of the world major languages have been developed and are being used but at present an OCR for Nastalique is not available and the published research on Nastalique OCR, Urdu OCR or even on any area of Urdu computing is almost non-existent, the reason being the challenges that the Nastalique style poses for its optical recognition. We used Matlab 7 for our experimentation the results are reported in this paper which are very encouraging.

Keywords
OCR, segmentation, Arabic script, histogram

1.0 INTRODUCTION

A single script with its basic character shapes is adapted for writing in multiple languages e.g. the Latin script for English, German, French etc. the Arabic script has been adapted for writing Persian, Urdu, Pashtu, Malay etc. Arabic writing has many forms and styles, the more common being Naskh, while its calligraphic counterpart is Nastalique, the most popular style of writing Urdu, the national language of Pakistan. Beautiful and decorative as it is, Nastalique is also highly cursive by nature.

The Urdu alphabet today contains 39 basic characters compared to Roman script languages which have a fewer number of characters in their alphabet. For this reason the development of an Urdu OCR is fairly more difficult than Roman script languages. So far no work has been done with regard to developing an Urdu OCR (Pal & Anirban, 2003)

Each of the Urdu characters has 2-4 different shapes according to their position in the word: isolated, initial, medial and final. Many character shapes have multiple instances. The shapes are context sensitive too – character shapes changing with changes in the antecedent character or the precedent one. At times even the 3rd or 4th character may cause a similar change depicting an n-gram model in a Markov chain (Sattar, Hyder & Pathan, 2006). Optical Character Recognition or OCR is the text recognition system that allows hard copies of written or printed text to be rendered into editable, soft copy versions. The problem of character recognition is the problem of automatic recognition of raster images as being letters, digits or some other symbol and it is like any other problem in computer vision (Parker, 1997).

2.0 LITERATURE SURVEY

A layout analysis for Urdu document image understanding is presented (Faisal, Adnan, Daniel & Thomas, 2006) highlighting its feature of right to left reading and writing order in contrast with Latin script languages that function from left to right. It considers the system of layout analysis as an important component of an OCR. The authors have experimented with a method of extracting text lines for image processing in the reading order of the Urdu script which is presented as an essential consideration for Urdu document image understanding.

A method is proposed for recognizing isolated characters claiming 98.3% accuracy for printed Urdu alphabet (Inam, Zaheer, Jehanzeb & Awais, 2007). However, the system also claims to be script independent and yet it is designed for Urdu only. The objective of the research is to develop an efficient recognition system for Urdu characters using minimum processing time.

A complete scheme for character recognition of totally unconstrained Arabic text based on a Model Discriminant HMM is presented (Alma'adeed, Higgens & Elliman, 2002). The system proposes feature extraction following the removal of variations in the word images which do not affect the identity of the written word. The system then encodes the skeleton and edges of the word and a classification process based on the HMM is used. The result is a word matching one in a dictionary. The study gives indication of successful results of a detailed experiment.

A word-level Arabic text recognition system is implemented (Erlandson, Trenkle & Vogt, 1996) that did not require character segmentation. They characterized the shape of Arabic words by unique feature vectors. These feature vectors were then matched against a database of feature vectors derived from a dictionary of known words. The database stored multiple

feature vectors for each word in a dictionary of 48,200 words. The word whose feature vectors strongly matched was returned as the hypothesis.

An Arabic word recognition system is designed and implemented (Albadr & Haralick, 1998) that recognized an input word by detecting a set of shape primitives on the word. The regions of words represented by these shape primitives were then matched with a set of symbol models. The description of the recognized word was obtained from a spatial arrangement of symbol models that were matched to regions of the word.

A holistic approach for Arabic word recognition was introduced (Khorsheed & Clocksin, 2000) which uses a normalization process to compensate for dilation and translation. The process adapted transforms the image of an Arabic word from Cartesian coordinates to polar coordinates similar to log polar transformation. Rotation is also converted into translation by this transformation.

A method was proposed (Bouslama & Kishibe, 1999) that combined the structural and statistical approach for feature extraction and a classification technique based on fuzzy logic. They segmented characters into a main and complement characters. The main segment was then centered and projected horizontally and vertically. The features of classification were then extracted from the number of complement characters and from the horizontal and vertical projection profiles of the main character. A set of fuzzy rules was used for classification. The recognition algorithm was tested on three different fonts and high recognition rates were achieved.

A structural probabilistic approach is presented (Amin & Mari, 1989) to recognize Arabic printed text. The system is based on character recognition and word recognition. Character recognition includes segmentation of words into characters using vertical projections and identification of characters. Word recognition is based on Viterbi algorithm and can handle some identification errors. The system was tested on just a few words and no figures were reported about its performance. The method has inherent ambiguity and deficiencies due to interconnectivity of Arabic text.

3.0 NASTALIQ SCRIPT CHARACTERSTICS

Nastalique script has the following characteristics:

- Text is written from right to left.
- Numbers are written from left to right.
- Urdu Nastalique script is inherently cursive in nature
- A ligature is formed by joining two or more characters cursively in a free flow form.
- A ligature is not necessarily a complete word, rather in most of the cases a part of a word, sometimes referred to as a sub-word.
- A word in Nastalique is composed of ligatures and isolated characters.
- Word forming in Nastalique is context sensitive i.e. characters in a ligature change shape depending upon their position and preceding or succeeding characters.

4.0 NASTALIQ SCRIPT SEGMENTATION

To segment a text region in an image into lines of text we use the horizontal projection profile (or histogram) of the text image.

Considering the image of size m×n is F (i, j), the projection of all foreground pixels perpendicular to the vertical axis and along the horizontal direction can be given as:

$$P_{\mathrm{h}}(i) = \sum_{j=1}^{n} F(i,j) \quad \text{for} \quad 1 \le i \le m$$

The horizontal projection profile of the text image separates the lines of text on the presence of white pixels between the two adjacent lines. A line of text covers the foreground pixels on the vertical scan from top of ascenders to the end of the descenders in the line of text, while scanning vertically from top to bottom.

One of the techniques for segmenting a text line image into ligatures and isolated characters is the vertical projection profile (or histogram) of the text line image, the projection of all foreground pixels in the image of extracted line L (i, j), having size r×s, perpendicular to the horizontal axis and along the vertical direction can be given as:

$$P_{v}(j) = \sum_{i=1}^{r} L(i,j) \quad \text{for} \quad 1 \le j \le s$$

The ligatures and the isolated characters present in the text line image are identified on the presence of white pixels separating them. The limitation of this technique is that it does not work in situations where adjacent ligatures overlap or shadow each other and the histogram so obtained portrays them as a single ligature.

5.0 EXPERIMENTATION AND RESULTS

Before the text image is given to the Nastaliq script segmentation system first it is binarized then the system segments the text image into lines of text and each of the lines of text is further segmented into ligatures and isolated characters.

Figure 1: Segmentation of text into lines

Figure 1 illustrates the separation of lines in the text image using horizontal projection profile or histogram; it also gives the correspondence between horizontal histograms and lines in the text image.

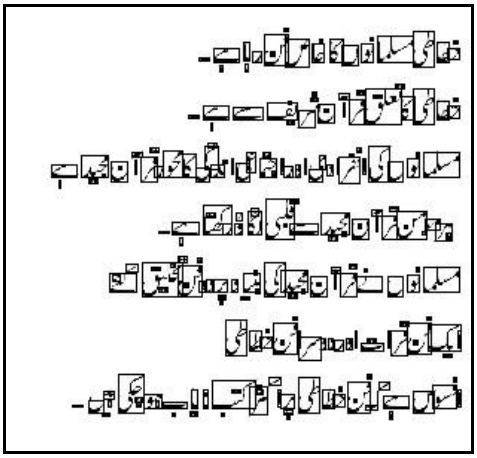

Figure 2: Seperation of ligatures and isolated characters

We used connected component labeling in Matlab to identify and isolate all the ligatures in the text image, illustrated in figure 2. This way we are able to isolate all ligatures and isolated characters. However, as an undesired effect all diacritic marks are also separated like dots belonging to certain characters. When a character has dots, this is identified with dots otherwise this will be interpreted as some other character. In Urdu we have different characters with the same base shape but having different numbers of dots, or dots at different locations or having no dots at all. The dots are to be associated with the character shape so that it can be interpreted as the correct character shape.

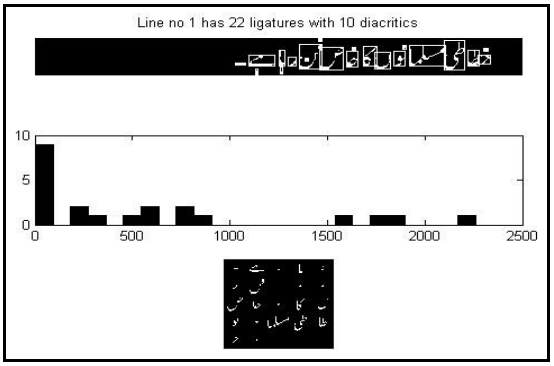

Figure 3: Analysis of text line 1

After segmenting a text image into separate lines, each of the text line images is assigned its number and processed separately further to identify the number of ligatures, isolated characters and diacritical marks it contains.

The given text image consists of seven lines (Figure 1). The text image is first segmented into seven lines and then each of the lines is further processed and the results are shown in figures 2 and 3.

The segmentation phase is shown in Figure 1, where the text image is first segmented into lines, and then each of the lines into ligatures, isolated characters and primitives. Here we call diacritic marks as primitives.

Figure 3 shows that it has 22 elements out of which 10 are diacritical marks and rest of the 12 are ligatures. The distinction between a ligature and a diacritical mark is made by trial and error method on counting the number of black pixels in each of the bounding box when connected component labeling is applied on separated text line images. By experimentation we found that if the number of black pixels in a bounding box is more than 25 then it is a ligature otherwise a diacritical mark.

In case of the Latin script or other discrete script OCRs we can use vertical projection profile (or histogram) to separate all characters in a line of text on the basis of small white spaces present between characters and their non-overlapping nature.

In figure 4 vertical projection profile is used on the vertical axis. Pixel density is plotted along horizontal axis (position on the text line). However, some ligatures could not be separated for recognition due to the overlapping between adjacent ligatures.

Once the text line images are separated from the text image and are processed as separate text image having only one line of text in the image.

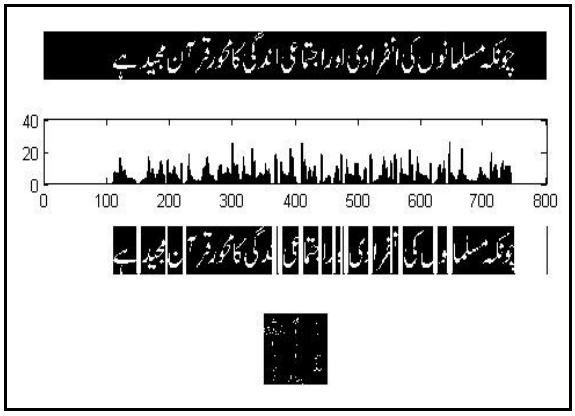

Figure 4: Ligature overlap line 1

In discrete scripts like Roman script for example individual characters are separated by the vertical projection profile of the text image with projecting all black pixels in the text line image along the vertical axis leaving white gaps in between the characters. In printed Roman text discrete characters do not overlap or shadow each other.

However in the case of Nastalique script, due to its cursiveness there is very much overlapping and shadowing of ligatures in the printed text, so ligatures cannot be separated by using a simple vertical histogram of text line image.

Figure 4 illustrate it clearly that there is much overlapping of words and ligatures in the printed Nastalique text and ligature separation is a more challenging job.

The text line image is scanned horizontally form left to right looking for the black (foreground) pixels and as they are found these are projected along the vertical axis to give the histogram showing distribution of pixel along the horizontal axis and the pixel densities along the vertical axis.

6.0 FUTURE WORK

In future we plan to use our Nastaliq script segmentation system to extend to a Nastaliq OCR system by training a learning machine to recognize the ligatures and isolated characters, like Neural Networks (NN), Support Vector Machines (SVM) or Hidden Markov Models (HMM).

REFERENCES

Albadr, B., Haralick, R. (1998). Segmentation-free approach to text recognition with application to Arabic text. *International Journal on Document Analysis and Recognition*, Vol. 1: 147-166.

Alma'adeed, S., Higgens, C. & Elliman, D. (2002). Recognition of Off-Line Handwritten Arabic Words Using Hidden Markov Model Approach, *Proc. 16th Int'l Conf. Pattern Recognition*, vol. 3, 481-484.

Amin, A., Mari, J. (1989). Machine recognition and correction of printed Arabic text. *IEEE Transactions on Man, Machine and Cybernetics*, 19(5): 1300-1306.

Bouslama, F., Kishibe, H. (1999). Fuzzy logic in the recognition of printed Arabic text. *IEEE Transactions on Pattern Recognition*, 1150-1154.

Erlandson, E., Trenkle, J., Vogt, R. (1996). Word level recognition of multifont Arabic text using a feature vector matching approach, *Proceedings of International Society for Optical Engineers,* SPIE, 2660: 63-70.

Faisal, S., Adnan, H., Daniel, K., & Thomas, M. (2006). Layout Analysis of Urdu Document Images. *Proceedings of 10th IEEE International Multitopic Conference, WMIC '06, Islamaad, Pakistan.*

Inam, S., Zaheer, A., Jehanzeb, K. & Awais, A. (2007). OCR For Printed Urdu Script Using Feed Forward Neural Network, *Proceedings of World Academy of Science, Engineering and Technology,* volume 23, ISSN 1307-6884.

Khorsheed, S., Clocksin, W. (2000). Spectral features for Arabic word recognition. *The IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP*, Istanbul, Turkey, 3574-3577.

Pal, U. and Anirban, S. (2003). Recognition for Printed Urdu Script, *Proceedings of the Seventh International Conference on Document Analysis and Recognition.*

Parker JR. (1997). Algorithms For Image Processing and Computer Vision, *John Wiley & Sons*.

Sattar, S. A., Hyder, S. S. & Pathan, M. K. (2006). Problems of Nastalique OCR: A comparison of Nastalique and Roman script OCRs, *Proceedings of the ICCCE* 06, Kuala Lumpur, Malaysia, Vol. 2, 1066-1071.