On the Segmentation of Multi-Font Printed Uygur Scripts

Anniwear YMIN† ‡ Yoshinao AOKI†
† Department of Information Engineering, Faculty of Engineering, Hokkaido University
13-8 Kita Ku, Sapporo-060, Japan
E-mail: eymhst@huie.hokudai.ac.jp
‡ Faculty of Electronics Information Science, Xinjiang University
Urumqi City, Xinjiang Uygur Autonomous Region, P.R. of China

Abstract

In many OCR systems, character segmentation is a necessary preprocessing step for character recognition. It is an important step because incorrectly segmented characters are not likely to be correctly recognized. The most difficult case in character segmentation is cursive scripts. Uygur character is a cursive script. This paper presents the problem of segmenting the Uygur characters in various fonts and size in printed scripts. The technique for the segmentation is presented as following: line separation, word separation, segmenting the word into isolated characters consists of the two step's algorithms, topological segmentation, and quasi-topological segmentation.

Topological segmentation is based on tracing the outer contour of a given word.

Quasi-topological segmentation is based on the decision to section a character on a combination of feature-extraction and character-width measurements.

Our approach relies on the feature of characters and fonts and profile models.

1. Introduction

Character segmentation is a technique that partitions images of lines or words into individual characters.

Character segmentation is fundamental to character recognition approaches which rely on isolated characters. It is a critical step because incorrectly segmented characters are not likely to be correctly recognized. The performance of any recognition system is affected by the presence of cursive scripts and the design of OCR systems must take the problem into account[1].

There are two main approaches for the recognition of cursive words. In the first approach termed "whole-word identification " the features of the entire word are used to identify the words without segmenting it. On the other hand the second approach, termed the "character-by-character identification " treats each word as a concatenation of a number of characters. Thus this method starts with the segmentation of each word into its characters using generally complicated Techniques. The second step involves recognizing each character and by using contextual information the word is identified[2]. In this paper, we used the second approach.

Uygur character is very similar to Arabic character. About Uygur character induced following section. Arabic character recognition is the most closely related to Uygur character recognition. Research in the area of Arabic character recognition did not start until the early 1980's. Some of the early work was done by Amin *etal.*[4], who used an on-line system called the interactive recognition of Arabic characters(IRAC). Badi and Shimura [5],[6] used the concept of contour tracing and identification of the component curves of the script for recognition. Masini [7] proposed an off-line system for multifont Arabic characters including segmentation of words into characters and identification of each character separately. A series of vertical and horizontal segments was extracted after scanning the character both vertically and horizontally. Almuallim[8] used a structural approach, where words are first segmented into strokes, and the geometrical and topological features of these strokes are then used to classify the words. Nouh *etal.*[9] used a standard set of Arabic characters to simplify the recognition process. Parhami[10] present a technique for the recognition of large font Farsi text of newspaper headlines. Al-Tikriti[11] introduces a fuzzy approach for some Arabic handwritten characters. El-Sheikh[12] employ Fourier descriptors to characterize outer contours of each character following

1015-4651/96 $5.00 © 1996 IEEE
Proceedings of ICPR '96

a segmentation process. El-Wakil[13] use a two stage hierarchical feature extraction scheme to represent and classify isolated handwritten Arabic characters. Mohamed Fakir and Chuichi Sodeyama[15] present a technique for the Arabic printed words is segmented into characters by analysing the vertical and the horizontal projection profiles using a threshold. In this paper, a new method for segmentation of cursive scripts is explained.

2. Characteristics of Uygur script

Uygur language is a Turkish language used in the Xinjian Uygur autonomous region in china and center Asia. Uygur script differs from Latin and Chinese characters in many structural features. Uygur character set is mostly in common with Arabic characters. A set of Uygur characters is shown in Fig.1. Uygur text is cursive in general. Uygur characters are normally connected on the writing line that is called "baseline". Uygur words are shown in Fig.2.

The Uygur language comprises 32 basic characters, Arabic language comprises 28 basic characters. The Uygur character is written from right to left. Most characters have four versions, depending on their position within a word, which is a group of joined or separated characters.

The four versions of characters are the head, middle, tail, and isolation(see Fig.3) types. The head version is always connected to another character at its tail (دەققەتسىز), the middle to characters at both sides (دەققەتسىز), and the tail to another at its start (دەققەتسىز) , otherwise they are regarded as separated versions (دەققەتسىز). Some characters contain more than one connected component: the main character body and one of four different stress marks shown in Fig.4. A stress mark may be over or under the character main body. It is a characteristic of the Uygur language that the main body of some characters is the same. For example, the three different characters (ز , ب , پ) have the same main body (ـر) but different stress marks. Unlike Chinese characters, Uygur characters are formed by loops, lines and curves. This makes it difficult globally to describe a character in one parametric form.

3. Preprocessing phase

Uygur texts were transferred to the computer through an image scanner. A preliminary processing consists of position normalization, baseline drift correction, line and word extraction.

	I	T	M	H		I	T	M	H
1	ا	ـا			17	ر	ـر		
2	ئ	ـئ			18	ز	ـز		
3	ۋ	ـۋ			19	د	ـد		
4	ى	ـى			20	ۇ	ـۇ		
5	و	ـو			21	ن	ـن		
6	ۆ	ـۆ			22	ق	ـق		
7	ۈ	ـۈ			23	ك	ـك		
8	ۇ	ـۇ			24	گ	ـگ		
9	ن	ـن			25	ڭ	ـڭ		
10	ت	ـت			26	غ	ـغ		
11	س	ـس			27			ـھ	
12	ش	ـش			28	ج	ـج		
13	م	ـم			29	چ	ـچ		
14	ل	ـل			30	خ	ـخ		
15	ب	ـب			31	ژ	ـژ		
16	پ	ـپ			32	ۋ	ـۋ		

Fig.1 The set of Uygur characters

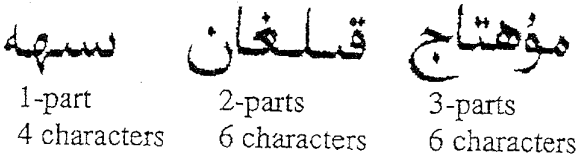


Fig.2 A selection of Uygur words

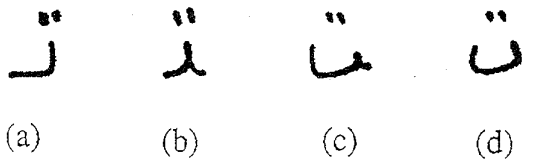


Fig.3 Different shapes of character "T" (a)Head,(b)Middle,(c)Tail,(d)Isolation.

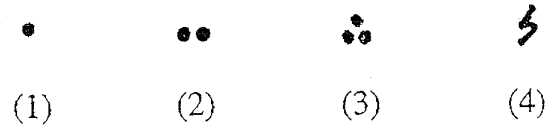


Fig.4 Different stress marks

3.1. Baseline drift correction

The image obtained from the text is converted to a binary matrix of "zeros"(white) and "ones"(black). The text to be recognized may be transfered to the system slanted. This fact affects the accuracy of the segmentation and the recognition. In this case baseline drift correction is needed. The hough transform is know as a method for the detection of straight lines in digital pictures by transforming each point of the image into sinusoidal curves. The Fourier transforms is know as a method also. By the way, a method using the Fourier transforms to correct the baseline drift for Chinese texts was reported basing on the power spectrum. This method is not very efficient, due to a long computing time. To solve this problem, a method based on the hough transform is proposed. In this paper, we used hough transform method to correct the baseline. Fig.5 shows an Uygur text before baseline drift correction, and Fig.6 shows an Uygur text after baseline drift.

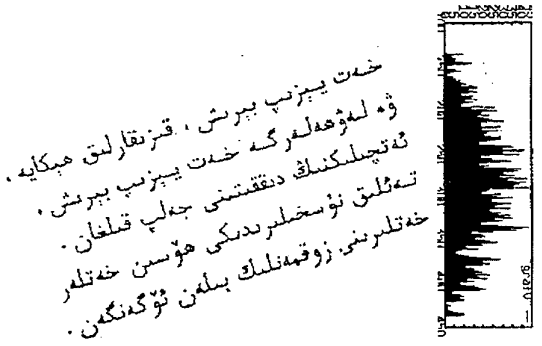


Fig.5 Before baseline drift correction

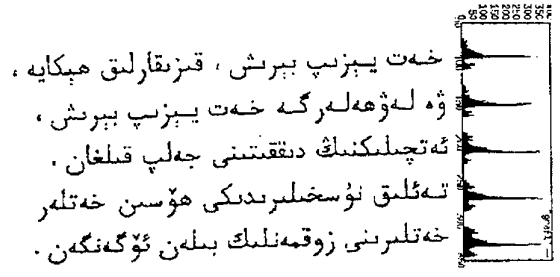


Fig.6 After baseline drift correction

3.2. Line and word extraction

After baseline drift correction, the next step transforms the sample into a single line of text by making a horizontal projection(equation 1).

$$f(i) = \sum_{j=0}^{n-1} p(i,j) \qquad i = 0, 1, 2, ..., m-1 \quad (1)$$

We use a fixed threshold to separate the pairs of consecutive line. Then, the line is segmented into words from a vertical projection. Figure 7 shown an extracted line and words.

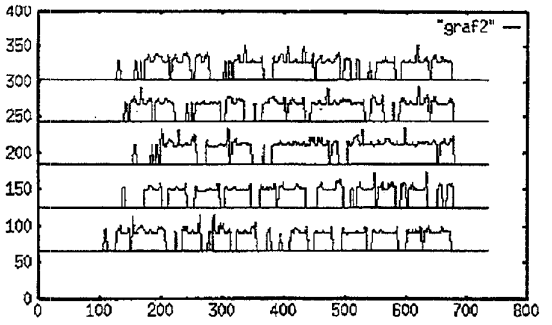


Fig.7 The line and word extraction

4. Segmentation phases

Each word can be easily isolated from the text since words are separated by enough space. On the other hand, segmenting the word into isolated character is a very difficult task especially for the Uygur language.

The segmentation phase is necessary in recognizing Uygur text. In this paper, it consists of three steps:(1) Segmentation of a word into zones.(2) Topological segmentation. (3) Quasi-topological segmentation.

4.1. The zones of word

This step is consists in segmenting a word vertically by horizontal lines into three zones that are denoted as the lower zone, the critical zone, and the upper zone(Fig.8).

The method consists in creating the horizontal projection of the word. The highest density value "max." is considered to be the baseline. A fixed "surface" threshold is used to decide the critical zone.

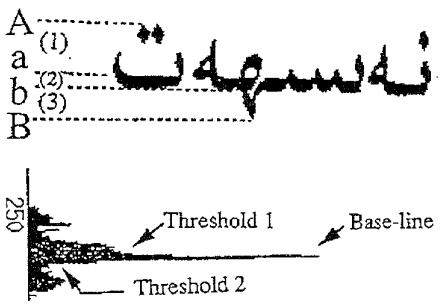


Fig.8 The main zones in a word.(1)Upper zone,(2)critical zone,(3)lower zone.

4.2. Topological segmentation

Topological segmentation (first segmentation) is based on the tracing the outer contour of a given word. From the edge of character strokes of upper zone the algorithm searches for the a possibility breaks along the vertical projection(Fig.9). The character formed by loop and other are segmented into two or over parts (Fig.10), Therefore, the first segmentation is rough, to solve this problem, The quasi-topological segmentation(second segmentation) method is proposed.

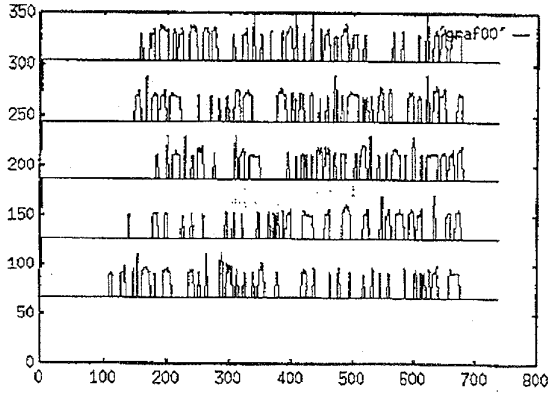


Fig.9 The horizontal projection profile of upper zone.

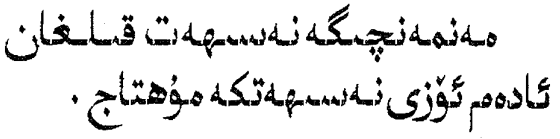


Fig.10 Result of first segmentation

4.3. Quasi-topological segmentation

The topological segmentation(first segmentation) involves to search a possibility breaks. The quasi-topological segmentation(second segmentation) bases the first segmentation, to section a character on a combination of feature-extraction and character-width measurements. The quasi-topological segmentation is a minutely step. Fig.11 shows after quasi-topological segmentation result. We used over two segmentation methods for segmenting the multi-font Uygur cursive script into characters. The Fig.12 shows results of other style segmentation.

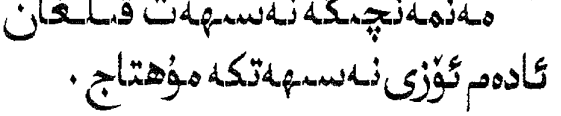
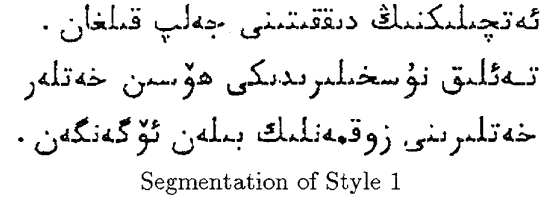


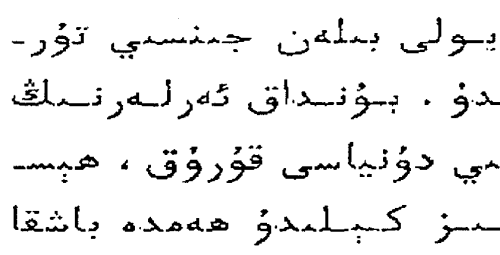
Fig.11 After quasi-topological segmentation



Segmentation of Style 1

5. Result

We present in this paper a method for the segmentation of Uygur Multi font scripts. We obtained a 93% average rate of success on Uygur Multi font words. Many of the segmentation errors are actually induced by poor-quality printers, or that the horizontal link between two consecutive characters is too short. Finally, in order to recognize the Uygur text at next step, we have to improve the method.



Segmentation of Style 2

Fig.12 The results of Uygur script segmentation

References

[1] R. L. Hoffman and J. W. Mccullough, "Segmentation Methods for Recognition of Michine-Printed Character," *IBM Journal of Research and Development* , pp. 153-165, March 1971.

[2] Talaat S El-Shelkh and Ramez M. Guindi,"Computer Recognition of Arabic Cursive Scripts," *Pattern Recognition*, Vol.21,No.4,pp.293-302,1988.

[3] Remain, "Machine recognition of handwritten Arabic words by airiest," in *Proc. 6th Int. Conf. Patt. Recogn.*, Oct.1982.

[4] A.Amin G.Masini, and J.P.Haton, "Recognition of Arabic words and sentences," in *Proc.7th Int. Conf. Patt.Recogn.*, 1984.

[5] K. Badi and M. Shimura, "Machine recognition of Arabic cursive scripts," in *Pattern Recognition in Practice.*, 1980.

[6] —-," Machine recognition of Arabic cursive script, " *Trans. Inst.Electron.Commun.Eng.*, vol. E65,no. 2, pp. 107-114, Feb.1982.

[7] A. Amin and G.Masini, "Machine recognition of multifont printed Arabic texts," in *Proc.8th Int.Conf.Patt.Recogn.*, 1986.

[8] H. Almuallim and S. Yamaguchi, " A method of recognition of Arabic cursive handwriting," *IEEE Trans. Patt. Anal. Machine Intell.*, vol.PAMI-9, no.5,Sept.1987.

[9] A.Nouh, A.Sultan, and R.Tolba, "On feature extraction end selection for Arabic character recognition," *Arab Gulf J. Scient.Res.*, vol. 2, no. 1, pp. 329-347, 1984.

[10] B. Parhami and M. Taraghi, " Automatic recognition of printed Farsi texts," *Patt.Recogn.*, vol. 14, nos. 1-6, pp. 395, 403, 1981.

[11] M. N.Al-Tikriti and S.K. Al-Ramchi, "Fuzzy approach for some Arabic handwritten character's computer recognition ," Computer Processing Arabic Language Workshop Papers,1985.

[12] T.El-Sheikh and R. Guindi, "Computer recognition of Arabic scripts," *Patt.Recogn.*, vol. 21, 1988.

[13] M. El-Wakil and A.Shoukry, "On line recognition of handwritten isolated Arabic charcters," *Patt. Recogn.*, vol.22,1989.

[14] H.Al-Yousefi and S.S.Udpa, "Recognition of Arabic Characters," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. 14, no. 8, 1992.

[15] Mohamed FAKIR and Chuichi SODEYAMA, 'Recognition of Arabic Printed Scripts by Dynamic Programming Matching Method," *Trans. IEICE* vol.E76-D, no. 2, Feb.1993.