An Effective Character Separation Method for Online Cursive Uyghur Handwriting

Mayire Ibrahim[1,3], Heng Zhang[2], Cheng-Lin Liu[2], and Askar Hamdulla[3]

[1] School of Electronic Information, Wuhan University, Wuhan 430072
mayire401@gmail.com
[2] National Laboratory of Pattern Recognition,
Institute of Automation of Chinese Academy of Science, Beijing 100190
{hzhang07,liucl}@nlpr.ia.ac.cn
[3] College of Software, Xinjiang University, Urumqi 830046
askar@xju.edu.cn

Abstract. There are many connected characters in cursive Uyghur handwriting, which makes the segmentation and recognition of Uyghur words very difficult. To enable large vocabulary Uyghur word recognition using character models, we propose a character separation method for over-segmentation in online cursive Uyghur handwriting. After removing delayed strokes from the handwritten words, potential breakpoints are detected from concavities and ligatures by temporal and shape analysis of the stroke trajectory. Our preliminary experiments on an online Uyghur word dataset demonstrate that the proposed method can give a high recall rate of segmentation point detection.

Keywords: online handwriting, Uyghur word recognition, character separation.

1 Introduction

With the widespread use of computing devices during the last decades, the need for fast and efficient text input measures is increasing. Handwriting recognition offers an important component in building such interfaces [1]. There has been considerable attention in the area of handwriting recognition for Latin-based and Oriental languages [2-5]. Despite that Uyghur script is widely used in regions of minority nationalities and the automatic recognition of handwritten Uyghur characters has many potential applications, the research on this problem has received little attention. This work considers online handwritten Uyghur word recognition and designs a character separation algorithm, which is an important step for word recognition.

There are two approaches for handwritten word recognition: global approach and analytical approach [6]. The global approach treats the word as a whole [7], while the analytical approach decomposes the word into smaller units or characters. Analytical approach is suitable for large word vocabularies because the number of character classes is limited. We adopt the analytical approach for handwritten Uyghur word recognition and need an effective algorithm for separating the characters in cursively written words. Many works have been done in handwriting segmentation [8-10] but

C.-L. Liu, C. Zhang, and L. Wang (Eds.): CCPR 2012, CCIS 321, pp. 530–538, 2012.

very few were for the segmentation of cursive Uyghur words [11]. To overcome the uncertainty of character segmentation before recognition, we adopt the strategy of over-segmentation for cursive word recognition: separate the word image into primitive segments, with each segment being a character or a part of character. After over-segmentation, correct characters can be formed by concatenating consecutive primitive segments. The segmentation of Uyghur words is difficult due to the characteristics of the language [12] and the variable styles of writing. Particularly, Uyghur characters have variable character sizes and gaps, and some characters have many variations of shapes depending on the position in words. In cursive writing, the characters in a word are mostly connected.

Based on the characteristics of cursive handwritten Uyghur words, our method first removes delayed strokes from the words, then potential breakpoints are detected from concavities and ligatures by temporal and shape analysis of the stroke trajectory. Afterwards, some redundant segmentation points are removed in a heuristic filtering step. Our preliminary experiments on online handwritten Uyghur words demonstrate that the proposed method can give a high recall rate of segmentation point detection.

The rest of this paper is organized as follows. Section 2 reviews the characteristics of Uyghur script. Section 3 describes the proposed character separation method. Section 4 presents the experimental results and Section 5 gives concluding remarks.

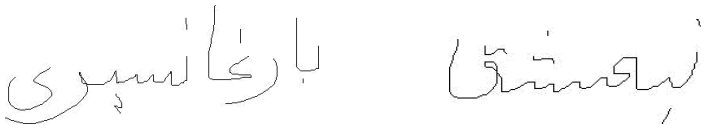
2 The Characteristics of Uyghur Script

Uygur language is a Turkish language used in the Xinjiang Uygur autonomous region in China. Uyghur writing is based on an alphabet and rules different from those of Chinese and Latin languages. Uyghur characters are written in a cursive style from right to left and no upper or lower case exists. Its alphabet contains 32 kinds of characters, and each character has two to four shapes and the choice of which shape to use depends on the position of the character (within its word or pseudo-word). Fig.1 shows some examples of Uyghur characters' shapes. Start form: only the suffix connects with the next character, Fig.1 (a); Middle form: initial and suffix connect with adjacent characters, Fig.1 (b); End form: only the initial connects with the above character, Fig.1 (c); Isolated form: initial and suffix does not connect with adjacent character, Fig.1 (d). Many characters have a similar shape. The position or number of these secondary strokes makes the only difference.

(a) (b) (c) (d)

Fig. 1. The variation of the Uyghur characters' shape: (a) Start form (b) Middle form (c) End form (d) Isolated form

The word consisting of a sequence of disjoint connected components is called pseudo-words. It has a main stroke that includes its basic shape, and complementary stokes which include dots or complementary parts. A Uyghur word can have one or more pseudo-words, shown in Fig.2. Each pseudo-word can be a group of characters or one character.



(a) (b)

Fig. 2. Examples of Uyghur words: (a) Words consisting of 5 pseudo-words; (b) Words with 7 characters connected

For character separation, we found that it's related with the writing style. The connecting types of different writing styles of Uyghur characters are as follows: ①Ligature: In order to make a word, all characters connect directly to the characters which immediately follow along a writing line or baseline. Some combination of two characters has special shapes called "ligatures". ② Concavity: Even for the same word, the different writers may have different writing style, resulting in different word shape such as concavity. ③ Overlap: Overlap refers to writing style of the points which have above or below writings in the same stroke. Characters in a word often overlap due to the writing styles.

3 Character Separation

As overview of the word segmentation system is shown in Fig3. Firstly, delayed strokes are removed from the online Uyghur word to release the segmentation difficulty. Then over-segmenting the word and filtering extra breakpoints using knowledge of character shapes.

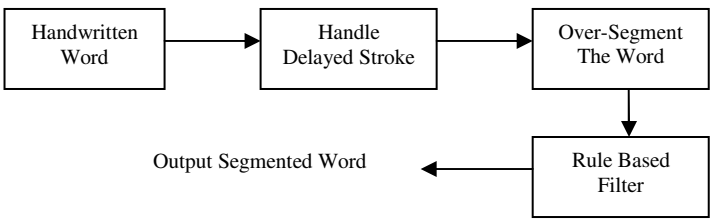


Fig. 3. An overview of segmentation system

3.1 Removing Delayed-Strokes

A stroke is defined as all-data point samples written between a certain pen-down action and the following pen-up action. Important information from the digitizing hardware other than the point coordinate pairs is the pen-down and the pen-up signals. Each group of strokes contain the primary strokes and secondary strokes which we call in this paper" delayed strokes", regardless the order by which they were written. Thus, a stroke may represent a pseudo-word or a character, or sometimes even a dot.

Detecting and removing delayed stroke is an important step. Delayed strokes are detected using a holistic approach. In order to examining the states of successively written Uygur strokes (either primary stroke or delayed stroke like dots for example), the following geometry features are calculated for each stroke such as width and height of bounding box and distance of overlapping between two strokes. If the value of the feature is less than predefined threshold, then the stroke is delayed-stroke. Fig.4 shows the original image and the one after removing the delayed strokes. After removing the delayed strokes, it is easier to segment the word.

(a) (b)

Fig. 4. (a) The original image; (b) Word image after removing delayed strokes

3.2 Over-Segmentation

In this section, after removing delayed strokes from the handwritten words, potential breakpoints are detected from concavities and ligatures by temporal and shape analysis of the stroke trajectory. Over segmentation is denoted when the character is segmented into several primitive segments. The over segmentation follows the steps below:

Step1: Formation of Initial Separation Point. For every right-to-left line $\overline{P_iP_{i+1}}$ in the stroke, P_i will be considered as an initial separation point between characters if the angle between $\overline{P_iP_{i+1}}$ and the horizontal axis is smaller than $\frac{\pi}{6}$.

Step2: Handling Space Overlapping. This stage will eliminate some of the elements in the group of initial separation points. If one of initial separation points has no any above writings across the whole range defined by the angle$(90-\alpha)^o$ to $(90+\alpha)^o$, then this point is accepted, else this point is rejected from the group of initial separation points (α is 25^oempirically), as shown in Fig.5.

(a) (b)

Fig. 5. (a) rejected segment point; (b) accepted segment point

Step3: Generating Separation Section. We compute the horizontal distance between every two initial separation points. If the distance is smaller than a predefined threshold (10 empirically), these points will form a bigger segmentation section. Some examples are shown in Fig.5 (a).

Step4: Locating Segmentation Points. In this stage our algorithm finds K possible separation sections in the main stroke. The middle of every separation sections will be located as segmentation points S_i (i=1, 2, ⋯, k), as shown in Fig.6 (b).

(a) (b)

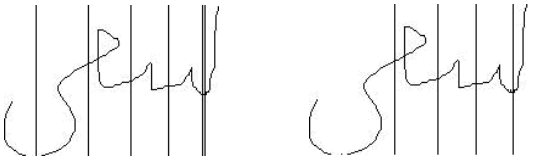
Fig. 6. (a) The result of step3; (b) The result of step4

3.3 Rule-Based Filtering

In this section, the suggested segmentation point is passed through the rule-based filtering to discard the incorrect segmentation point.

Rule1: baseline is computed with the respect to the horizontal pixel density. The baseline corresponds to the original writing line in which all the connection between the successive characters take place. For handwriting, the base line is an ideal concept and simplification of actual writing. In practice connections occur near, but not necessarily on a baseline. Compute the vertical distance D between suggested segment point S_i and the y-value of the baseline on a word. If the distance is less than a predefined threshold (10 empirically), then filter the segmentation point. An example is shown below in Fig7.

Rule2: If the distance between two suggested segmentation points is less than a predefined threshold (5 empirically), remove the segmentation points.



(a) (b)

Fig. 7. Filtering example: (a) Segment results before applying rule; (b) Segment results after applying rule

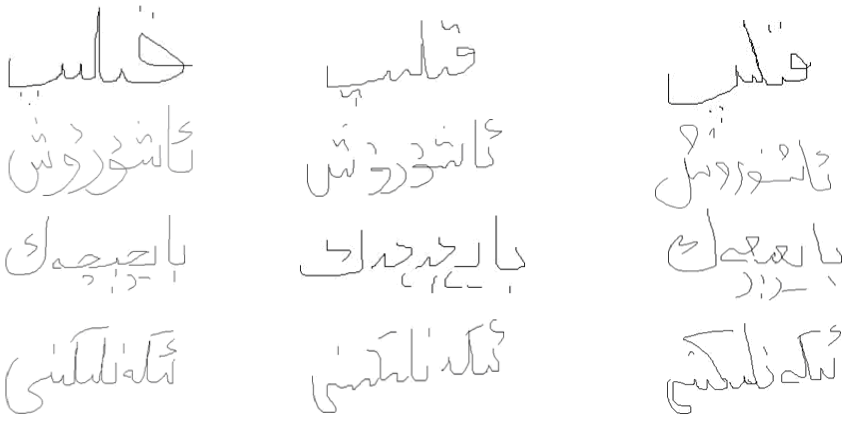
4 Experimental Results

4.1 Acquisition of Online Handwritten Word Data

Acquisition of online handwriting data is the first step of Uyghur online handwritten recognition system. As we know, there is no publicly available online Uyghur database. In this paper, the database we used for testing consists of 900 words collected from different people (300 words written by each writer) using Han Wang writing tablets. The handwriting input is captured as a stream of positions in the form of "x" and "y" coordinates. Depending on the type of a digitizer it may be able to provide more information such as pen-pressure and pen-tilt and with programming one can also compute the pen movement speed. Most systems including ours, however, use only the coordinates and pen-up/down signals. The information of the word image is saved as a format of binary file.

Table 1. The number of test word (one set)

Number of characters (N)	$2 \leqq N \leqq 5$	$6 \leqq N \leqq 10$	$11 \leqq N \leqq 18$	Total Word	Total character
Number of test word	65	203	32	300	1898



(a) set1 (b) set2 (c) set3

Fig. 8. samples of different three set data

The number of characters of in a word varies from 2 to 15 as summarized in table 1. Set 1 is written in a more regularly style and set 2 is written in a normal while the third is written more free. A few samples of different three set data are shown in Fig8.

The algorithms were implemented in C++ in platform of Microsoft visual C++ 6.0.

4.2 Experimental Results

To examine the utility of our approach, we have carried out experiments on the word data sets explained in the previous section. For the small sample test, the character segmentation algorithm is usually evaluated with the help of manual statistics; for the large sample image test, the final recognition accuracy of the system applied is often used to evaluate the performance of the segmentation algorithm. In this paper, the first method is adopted. We evaluate the performance of character detection in terms of the rates of Recall (R) and Precision (P), which are defined as:

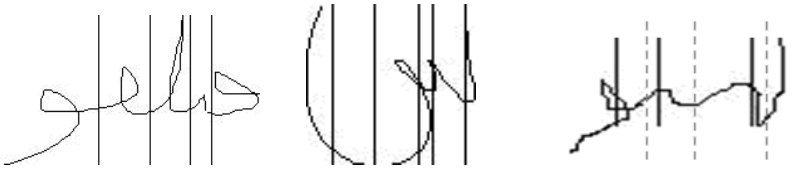
$$R=\frac{\text{number of correctly detected seperation point}}{\text{number of true seperation point}} *100\%$$

$$P=\frac{\text{number of correctly detected seperation point}}{\text{number of detected seperation point}} *100\%$$

Table 2 summarized the recall and precision of our approach on different sets for online handwritten Uyghur word. The maximum number of suggested separation points exceeds the right separation points by 40%-50%. A few examples of segmentation result are shown in Fig9. In Fig9.(a) shows one word with 5 characters connected. This word segmented into 5 characters correctly. There is no any

Table 2. Performance of online handwritten Uyghur word separating on different dataset

	Total successful segment point	Total over-segment point	Recall (%)	Precision (%)
Set1	1875	3298	98.78	56.19
Set2	1865	4694	98.26	52.87
Set3	1832	3465	96.52	49.73



(a) (b) (c)

Fig. 9. (a) Examples of correct segment results; (b) Examples of over-segment results; (c) Examples of incorrect segment results

over-segment point or incorrect segment point. In Fig9.(b) shows one isolated character in which have 5 over-segment point. Actually, there is should not have any segment point in the character. In Fig9.(c) shows one main stroke with 4 characters. There is have 3 incorrect segment points in this main stroke, because writing style of the one who have written this word is so free. The main stroke should have 3 segment points with the dotted line indicates.

5 Conclusions

This paper presented a new approach for character separation of online Uyghur cursive handwriting which can lay the foundation for the word recognition task. The cursive nature of Uyghur, delayed strokes and characters overlap are some of the key problems that make Uyghur word segmentation more difficult than other languages such as Latin or Chinese. Our preliminary experiments on Uyghur word data demonstrate that the proposed method can give a high recall of segmentation point detection. The proposal for future work is to refine the filtering of separating point for improving the precision and to apply a recognition system for testing the separating performance.

Acknowledgments. This work is supported by Program for New Century Excellent Talents in University (NCET-10-0969) and Natural Science Foundation of China (No.61163033, 60825301, 60933010).

References

[1] Plamondon, R., Srihari, S.N.: On-line and off-line handwriting recognition - a comprehensive survey. IEEE Trans. Pattern Anal. Mach. Intell. 22(1), 63–85 (2000)
[2] Jaeger, S., Manke, S., Reichert, J., Waibel, A.: Online hand-writing recognition: The NPen++ recognizer. Int. J. Document Analysis and Recognition 3(3), 169–180 (2001)
[3] Khorsheed, M.S.: Off-line Arabic character recognition - a review. Pattern Analysis and Applications 5(1), 31–45 (2002)
[4] Liu, C.L., Koga, M., Fujisawa, H.: Lexicon-driven segmentation and recognition of handwritten character strings for Japanese address reading. IEEE Trans. Pattern Anal. Mach. Intell. 24(11), 1425–1437 (2002)
[5] Wang, Q.F., Yin, F., Liu, C.L.: Handwritten Chinese text recognition by integrating multiple contexts. IEEE Trans. Pattern Anal. Mach. Intell. (in press, 2012)
[6] Amin, A.: Off-line Arabic character recognition: the state of art. Pattern Recognition 31(5), 517–530 (1998)
[7] Benouareth, A., Ennaji, A., Sellami, M.: Arabic handwritten word recognition using HMMs with explicit state duration. Journal on Advances in Signal Processing, 1–13 (2008)
[8] Casey, R.G., Lecolinet, E.: A survey of methods and strategies in character segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 18(7), 690–706 (1996)
[9] Cheung, A., Bennamoun, M., Bergmann, N.W.: A new word segmentation algorithm for Arabic script. DICTA: Digital Imaging Comput. Tech. Appl., 431–435 (1997)

[10] Lu, Y., Sridhar, M.: Character segmentation in handwritten words: an overview. Pattern Recognition 29(1), 77–96 (1996)
[11] Halmurat, A.: Research and development of a multifont printed Uyghur character recognition system. Chinese Journal of Computers 27(11), 1480–1484 (2004) (in Chinese)
[12] Sadik, M.: Basics of Uyghur Language. Xinjiang People's Press, Urumqi (1992) (in Chinese)
[13] Abdelazeem, S., Eraqi, H.M.: On-line Arabic handwritten personal names recognition system based on HMM. In: Proc. 11th Int. International Conference on Document Analysis and Recognition (ICDAR), Beijing, China, pp. 1304–1318 (2011)