H-014

A Development of Pattern-based Online Handwriting Uyghur Character Recognition System

Yidayet Zaydun, Tsuyoshi Saitoh

Computer Graphics Laboratory, Tokyo Denki University
2- 2, Kanda Nishikicho, Chiyoda-ku, Tokyo 101-8457, Japan
E- mail: yidayet@cgl.im.dendai.ac.jp, saitoh@im.dendai.ac.jp

Abstract

This paper describes the experimental results of pattern-based online Uyghur character recognition. For the purpose of classification of the sub-word patterns, samples containing 4 million characters were collected and analyzed. As the result of the analysis, we found that a pattern-based recognition approach was applicable effectively, because average recognition rate on the 12 volunteers' data became 91% for one-character sub-words and 87% for two-character sub-words.

We also discuss that a better recognition rate will be obtain by improving the standard pattern database for our system.

1. Introduction

The Uyghur language is traditionally used a modified Perso-Arabic alphabet known as Chagatai script since the 10th century. A further modified Arabic script with additional diacritics to distinguish Uyghur vowels was introduced in 1983 and is being used now. Uyghur script is a cursive style and written from right to left, similar to Arabic. But it is different from Arabic in some terms. In Arabic, a normal text is composed only of series of consonants; thus, the word *shukran*, "Thank you", is written شكراً. In Uyghur, the vowels are mandatory. So, this word is written شۇكران. This complicates Uygur character recognition than Arabian character recognition. Both of short and long Arabic vowels are used in full-vocalized text. However, it is optional and usually is reduced in actual writing, especially on the web. The number of vowels and its usage is a special feature of the Uyghur language. It makes a great difference between the Uyghur and the Arabic language, also makes it difficult to recognition.

Many works have been done on the recognition of Latin and Arabic characters, both of online and off-line, but very little has been done on Uyghur. Moreover, the research on the online Uyghur character recognition is much less than off-line recognition.

Up to now, no concrete result has been accomplished yet though some researchers are studying on Uyghur character recognition, especially in off-line recognition [1]. Although an off-line document recognition system was proposed and developed by the researchers of Xinjiang University and Tsinghua University in China, the system had several problems such as low recognition rate and quite low speed.

In this research, we aim at the online recognition of Uyghur characters written with a graphic tablet, and develop the recognition system as a user interface for the portable digital devices such as PDA and cellular phones.

2. Sub-word

A word contains one or more basic isolated blocks, which defined as sub-word. A sub-word is constructed of one or more shapes. Each sub-word is separated from others by a small space. The text can be easily resolved into sub-words in this respect.

The Uyghur language is constructed of 126 different shapes of 32 basic characters. Each character contains more than two different shapes: initial, medial, final and isolated. Some characters are written by only one stroke, but many characters are composed of two parts, primary stroke and secondary stroke [2], and some characters differ by secondary stroke(s) but the primary stroke is exactly the same as shown in Fig 1.

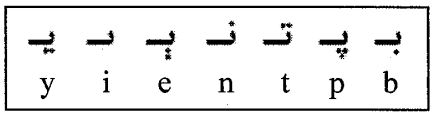

Figure1. Characters differ by secondary stroke

3. Pattern-based recognition

Most Uyghur characters are composed by less than three strokes. That is to say, the Uyghur writing is almost writer-independent even for left-handed person. Usually, multi-character sub-words needs segmentation into single characters in character-based recognition. Also, stroke number/order of a multi-character sub-word causes an direct influence on recognition speed and recognition rate. For the purpose of developing a none-segmentation and stroke-number-free or stroke-order-free recognition system, the pattern-based recognition is attracting our attention. The base of our opinion is that, many characters and sub-words differ only in secondary strokes but their primary stroke is almost the same. For example, the two-character sub-word pattern "بو" has 31 different sub-words in actual writing shown in Fig2. All these sub-words can be classified only one pattern based on their primary stroke.

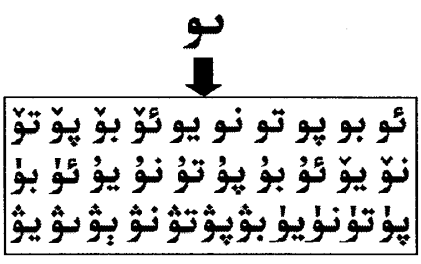

Figure2. Sample of a sub-word pattern

3.1 Sub-word pattern analysis

We defined 16 patterns for isolated shape. That is to say, there are 16 kinds of patterns for one-character sub-words because it constructed by isolated shape only. The two-character sub-word is constructed by an initial shape and a final shape. Uyghur alphabet has 22 initial shapes and 33 final shapes. Thus, totally 726 kinds of two-character sub-words can be theoretically combined. We defined 8 patterns for initial shape and 16 patterns for final shape. Thus, totally 128 kinds of patterns is possible to combine.

For the purpose of understanding the pattern features of Uyghur writing, we collected sample texts consisting of 4,140,413 characters and held statistics of patterns for the sub-words consisting of one-character to four-character. The result of one-character/two-character sub-word patterns is shown in Table 1.

Table1. Pattern in one-character/two-character sub-words

Sub-word type	One	Two
Number of sub-words	541,590	539,702
Combination of sub-words	16	726
Sub-words appears actually	16	305
Combination of patterns	16	128
Patterns appears actually	16	79

The result shows that, it is possible to recognize one-character and two-character sub-words by the recognition of approximately 320 sub-words or around 100 patterns actually appeared in actual writing. That is to say, the recognition processing will be decreased by nearly 40%.

3.2 Combination

Based on above discussion, we implemented a combination process for our recognition system. In recognition process, the input sub-word is dividing into primary stroke and one or several secondary strokes (when secondary strokes are available), and would be recognized. In the combination process, the first secondary stroke will be combined with the primary stroke and output a combined sub-word as a result. Then, the second stroke will be combined with the combination result. This process continues until the last secondary stroke is combined. The final result will be output as a recognized sub-word. Thus, it is possible to give the same result by the combination processing even when the stroke order is different.

3.3 Pattern-based recognition

In the purpose of develop a stroke-order-free recognition system such as [3], we implemented a pattern-based recognition process with combination process. We applied the method of Approximate Stroke Sequence String Matching (ASSSM) [4]. It is simple and well-understood method and easy to implement with the quite small database.

The recognition and combination process tested on a sample of freely-written 12 data sets. Each data set includes 94 one-character sub-words and 170 two-character sub-words. The recognition results of one-character and two-character sub-words are showed in Table 2. Pattern database used in this experiment includes 112 patterns of primary stroke and 27 patterns of secondary stroke, including some different stroke sequence of the same sub-word pattern.

Table2. Recognition result

Sub-word	N-best accumulative recognition rate(%)				
	1	~2	~3	~5	~10
One-char	91.40	97.25	98.94	99.73	99.73
Two-char	88.55	94.68	97.55	98.66	99.18

The table showed that, almost all sub-words recognized in best three candidates by pattern-based recognition.

4. Conclusions

We have proposed pattern-based recognition method for online handwriting Uyghur character recognition. Experimental results showed that, it is effective to recognize one-character and two-character sub-words.

Based on these results we considered that, the pattern-based recognition method is more effective to recognize other sub-words; for three-character sub-words, it is possible to recognize all 22×22×33=15,972 sub-words by the recognition of 8×8×16=1024 patterns. Besides, a lot of patterns are never used in actual writings. Thus, the recognition at shorter time is enabled and a higher recognition rate can be obtained.

The testing and improving of pattern-based recognition method for other sub-words is one of the targets of our future works.

References

[1] A. Ymin, Y. Aoki: "On the Segmentation of Multi-Font Printed Uyghur Scripts", in Proc. of 13th Int. Conf. on Pattern Recognition (ICPR'96), Vol.3, pp.215-219 (1996).

[2] Y. Zaydun, T. Saitoh: "Uyghur Character Recognition using the Imaginary Stroke Information", in Proc 2007 IEICE General Conf., pp. 255 (2007).

[3] M. Nakai, H. Shimodaira and Sh. Sagayama: "Generation of Hierarchical Dictionary for Stroke-order Free Kanji Handwriting Recognition Based on Substroke HMM", in Proc. of Int. Conf. on Document Analysis and Recognition (ICDAR2003), Vol.1, pp. 514- 518 (2003-08).

[4] S. H. Cha, Y. C. Shin and S. N. Srihari, Approximate Stroke Sequence String Matching Algorithm for Character Recognition and Analysis. 5th ICDAR, Banglore, India, 1999.