A Pattern-based Recognition Method in Online Handwriting Uyghur Character Recognition

Yidayet Zaydun[1], Tsuyoshi Saitoh[1]

[1] Computer Graphics Lab., Tokyo Denki University
2-2 Kanda Nishikicho, Chiyoda-ku, Tokyo 101-8457, Japan
yidayet@cgl.im.dendai.ac.jp, saitoh@im.dendai.ac.jp

Abstract

This paper discusses pattern-based online handwriting Uyghur character recognition. A statistical analysis result showed that the sub-word is the basic unit of Uyghur writing. As a first step of the implementation of a sub-word-based recognition system, the one-character sub-words were recognized using Approximate Stroke Sequence String Matching method, and got a recognition rate about 97%. Experiment result proved this method also effective to the recognition of other sub-words. To get a clear idea about sub-word patterns, the sample of containing 4 million characters was collected and analyzed. Based on the analysis result, a pattern-based sub-word recognition approach is suggested. The recognition result of one-character sub-words is indicated.

We discuss that it is possible to recognize 726 kinds of two-character sub-words by the recognizing half of their patterns only. In addition, it is believed that an ideal recognition rate can be obtaining by combining other features such as sub-word frequency. The recognition of other sub-words is the targets of our future work.

Keywords: Pattern recognition, Online recognition, Handwriting Uyghur character

1. INTRODUCTION

Many works has been done on the recognition of Latin, Chinese, Japanese and Arabic characters, both of online and off-line, but very little has been done on Uyghur. It is a difficult task because of the cursive writing and the different sizes of different characters. Also, it is true that the research of the online recognition is much less than off-line recognition.

There is no concrete result has been accomplished yet though some researchers are studying on Uyghur character recognition, specially in off-line recognition [1][2]. Even an off-line document recognition system was proposed and developed by the researchers of Xinjiang University and Tsinghua University in China, but it is said that the system was not able to be achieved at the actually applying level because of the low recognition rate and various unsolved problems.

In this research, it aims at the online recognition of Uyghur characters written with pen tablet, and develops the recognition system as a user interface for the portable digital devices; finally makes Uyghur script can be used freely on the every digital product in Uyghur society.

In this paper, the sub-word structure of Uyghur script is explained; our previous works and its results are presented, sub-word patterns and its frequency in real writing are analyzed, and a pattern-based recognition system is suggested.

2. ABOUT UYGHUR LANGUAGE

The Uyghur (also spelled as Uighur) is an ethnic group who lives primarily in northwestern China. They have their own language and script. Uyghur language belongs to the southeastern branch of the Turkic language family. It is spoken by over 10 million Uyghur people in the world.

2.1 Uyghur script

The language traditionally used a modified Perso-Arabic alphabet, known as Chagatai script since the 10th century. A further modification of the Arabic script, with additional diacritics to distinguish Uyghur vowels, was introduced in 1983 and is being used now.

Uyghur language has 32 basic characters in the alphabet. Each character contains more than two different shapes. They are: initial, medial, final and isolated. The initial shape is connected to others at its tail, the medial shape is at both sides, and the final shape is to others at its start; otherwise they are regarded as isolated shape.

2.2 Primary stroke and secondary stroke

Uyghur text is constructed by 126 different shapes of 32 basic characters. Some characters are written by one stroke only, but many characters are composed of two parts, primary stroke and secondary stroke [3], and some characters only differ by secondary stroke(s) but the primary stroke is exactly the same, shown as Fig1. Using a combination of these secondary strokes above and below of these primary strokes, the full complement of 32 characters can be constructed.

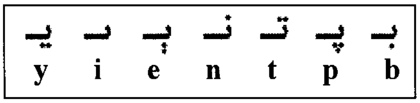

Fig1. Characters differ by secondary stroke

2.3 Sub-word

Uyghur script is a cursive script and written from right to left. The sizes of characters differ from character to character and from one shape to another, making recognition techniques difficult. However, a word contains one or more basic isolated blocks. We defined these blocks as sub-word. A sub-word constructed by one or several shapes. Each sub-word is separated from others by a small space. The text can be easily resolved to sub-words by this respect. The sample text which means "Uyghur Language" including 2 words, 4 sub-words and 9 characters, is shown in Fig2.

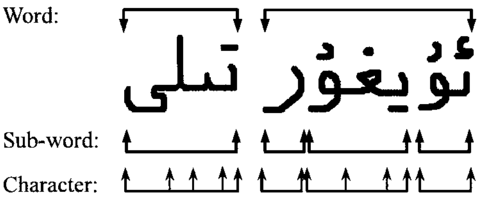

Fig2. Word and sub-word

3. PREVIOUS WORK

A feature of our former work is that the various shapes of one character are considered as different characters. The characters chosen for the experiment are the whole 126 different shapes of 32 basic characters.

3.1 Sub-word structure analysis

We held a sub-word analysis on a sample text consists of about 1.4 million characters. Analysis result showed in Table1.

Table1. Sub-word statistics

Characters per sub-word	Number of sub-words	%
1	267682	40.803
2	207183	31.581
3	91112	13.888
4	42016	6.405
5	25959	3.957
6	11470	1.748
7	6740	1.027
8	2315	0.353
9	1056	0.161
>=10	506	0.077
Total	656039	100

This sample consists of 656039 sub-words. The average word length is 6.25 characters or 2.92 sub-words, and average sub-word length is 2.13 characters. Based on this result we discuss that, the basic block to be dealt with in Uyghur character recognition system should be the sub-word rather than the word.

Sub-words can be classified into 3 groups according to the numbers of the characters they contain. First group is the sub-words consisting of only one isolated shape. Hence, it is not necessary to segmentation. Second group is the sub-words consisting of two shapes. It needs segmentation in two parts only. Last group contains of all other sub-words, and it needs segmentation in three or more parts [4].

3.2 Recognition method

Various recognition methods are used in online character recognitions. The adopted method in this research is that converting the input data to sequence string, and comparing it to the standard character models in database by the method of Approximate Stroke Sequence String Matching (ASSSM) [5].

In first step, the stroke sequences of input data is extracting into integer array based on the 8-direction stroke convention, shown as Fig3. The distance (which means disparity) between two strokes would be computed by ASSSM after the stroke sequence array is obtained. Input character would be compared to every referencing characters in sequence string database and best compatible character (which one that has the shortest distance) could be found. In other words, the input character is recognized.

$$\begin{pmatrix} \nwarrow & \uparrow & \nearrow \\ \leftarrow & & \rightarrow \\ \swarrow & \downarrow & \searrow \end{pmatrix} = \begin{pmatrix} 3 & 2 & 1 \\ 4 & & 0 \\ 5 & 6 & 7 \end{pmatrix}$$

Fig3. The 8-direction stroke convention

3.3 Recognition rate

We get a recognition rate about 97% in average on a sample of freely-written 8 data sets, showed in Table2 as N-best accumulative recognition rate [6].

Table2. Recognition rate

	N-best accumulative recognition rate				
	1	~2	~3	~5	~10
Rate(%)	91.35	96.59	97.7	98.17	99.21

Recognition rate was improved by using the imaginary stroke information [3], recognition with character separation based on character structure [7], and improving sampling algorithm.

It is believed that an ideal recognition rate can be obtain while using other characteristics such as word/sub-word frequency, pen pressure and other useful information in handwritings.

4. PATTERN-BASED RECOGNITION

In the purpose of develop a none-segmentation and stroke-number-free or stroke-order-free recognition method, the pattern-based recognition is attracting our attention. the base of our opinion is that, many characters and sub-words only differ by secondary strokes but their primary stroke is almost the same (Fig1). The patterns can be classified based on the primary strokes in the sub-word.

4.1 Sub-word patterns

We defined 16 patterns in one-character sub-word, showed in Table3. The patterns of initial, medial and final shapes are showed in Table4. The initial shape patterns are 8 and the final shape patterns are 16. It is said that, the patterns of two-character sub-words are 8×16=128 at most. Also, the patterns of medial shape are 8, and it makes 1024 patterns for three-character sub-words at most.

Table3. One-character sub-word patterns

ا	ا	ٯ	ف ق
ە	ە	ل	ك ڭ ل
ٮ	ب پ ت ن	ک	گ
ح	ج چ خ	م	م
د	د	ھ	ھ
ر	ر ز ژ	و	و ۇ ۆ ۈ ۋ
س	س ش	ى	ې ى ي
ع	غ	لا	لا

Table4. Patterns of initial, medial and final shapes

(a) Initial Shape

ٮـ	ئـ بـ پـ تـ نـ ېـ ٮـ يـ
حـ	جـ چـ خـ
سـ	سـ شـ
عـ	غـ
ڡـ	فـ قـ
لـ	کـ گـ ڭـ لـ
مـ	مـ
هـ	هـ

(b) Medial Shape

ـٮـ	ـئـ ـبـ ـپـ ـتـ ـنـ ـېـ ـٮـ ـيـ
ـحـ	ـجـ ـچـ ـخـ
ـسـ	ـسـ ـشـ
ـعـ	ـغـ
ـڡـ	ـفـ ـقـ
ـلـ	ـکـ ـگـ ـڭـ ـلـ
ـمـ	ـمـ
ـهـ	ـهـ

(c) Final Shape

ـا	ـا	ـى	ـي ـى ـې
ـە	ـە	ـح	ـج ـچ ـخ
ـٮ	ـب ـپ ـت ـن	ـس	ـس ـش
ـد	ـد	ـع	ـغ
ـر	ـر ـز ـژ	ـٯ	ـف ـق
ـک	ـگ	ـل	ـك ـڭ ـل
ـو	ـو ـۇ ـۆ ـۈ ـۋ	ـم	ـم
ـلا	ـلا	ـه	ـه

4.2 Pattern analysis

In the purpose of understand the pattern features of Uyghur writing, we collect a sample text consists of 4,140,413 characters and held a statistics of patterns for the sub-words consists of one-character to four-character. The result of one-character sub-word patterns is showed in Table5.

Table5. Pattern in one-character sub-words

Pattern	Frequency	Pattern	Frequency
ر	176040	ٯ	12349
و	108227	س	8178
ا	62780	م	4148
ه	50194	ح	1598
د	43114	لا	722
ٮ	37608	ھ	309
ل	19169	ع	263
ى	16741	ک	150

The results showed that, the total isolated characters in this sample are 541,590, makes about 13.06% of total characters. Analysis result also showed that, it is possible to recognize all one-character sub-words by the recognition of 16 patterns. It meaning that, the recognition processing will be decrease by about 50% and the recognition time can be shortening.

The analysis results of the sub-words constructed by two characters are showed in Table6.

Table6. Pattern in two-character sub-words

Total characters	4,140,413
Sub-words by two-characters	539,702
Possible combinations of sub-word	726
Sub-words appears actually	305
Possible combinations of pattern	128
Patterns appears actually	79

The results showed that, it is possible to recognize all two-character sub-words by the recognition of about 300 sub-words or only 80 patterns actually appeared in real writing. It is said that, the recognition processing will be decreasing by about 37%.

The results of three-character sub-words also showed that, it is possible to recognize 1676 actually appeared three-character sub-words by the recognition of about 320 patterns (31% of total) only.

Based on these results we discuss that, it is possible to recognize from the one-character sub-words to the three-character sub-words by using a pattern database consists of about 420 patterns.

In next step, the combination rule is needed for the processing that combines the recognized pattern with the secondary strokes and makes recognized sub-word outputs.

4.3 Combination rules

Based on above discussion, we implemented a combination process for our recognition system. In recognition process, the input character is dividing into primary stroke (PS) and secondary stroke (SS), and recognized respectively. The final recognized character (or sub-word) can be combined in combination process.

4.4 Recognition system

In the purpose of develop a stroke-order-free recognition system such as [8], we implemented a pattern-based recognition process with combination rules, showed in Fig4.

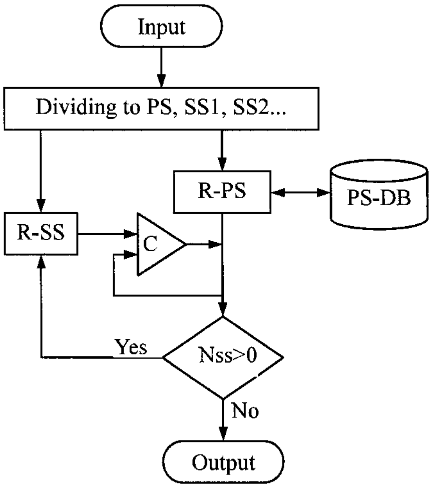

Fig4. Recognition system

here:

PS: Primary stroke

SS: Secondary stroke

R-PS: Recognition result of primary stroke
R-SS: Recognition result of secondary stroke
PS-DB: Pattern database of primary stroke
Nss: Number of secondary stroke
C: combination process

In this process, the longest stroke in input data is considered as primary stroke and others as secondary stroke. The number of secondary stroke will be checked after primary stroke is recognized, or combination process is finished. The process will finish and brings a recognized output character or sub-word when the number of secondary stroke equal to 0. Thus, it is possible to give the same result by the combination processing even when the stroke order is different.

4.3 Experiment

The recognition and combination process tested on a sample of freely-written 8 data sets. The recognition results of one-character sub-word are showed in Table7.

Table7. Recognition result with combination

Data set	Previous work	Pattern recognition
T01	84.85%	93.94%
T02	84.85%	96.97%
T03	84.85%	90.91%
T04	84.85%	81.82%
T05	84.85%	93.94%
T06	75.76%	81.82%
T07	78.79%	84.85%
T08	78.79%	84.85%
Average	82.20%	88.64%

The table showed that, our process is effective to one-character sub-word patterns: The average recognition rate is improved about 6.44%.

5. CONCLUSION

We have proposed pattern-based recognition method for online handwriting Uyghur character recognition. Experimental results showed that, it is able to obtain higher recognition rates.

It is believed that, the pattern-based recognition method is more effective to recognize other sub-words; for two-character sub-words, it is possible to recognize all 22×33=726 sub-words by the recognition of 8×16=128 patterns. Besides, a lot of patterns are never used in real writing. This not only makes the recognition at shorter time possible but also a higher recognition rate can be obtained.

The testing of pattern-based recognition of 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] Halmurat, Arzigul: "Research and Development of a Multifont Printed Uyghur Character Recognition System", Chinese Journal of Computers, Vol.27, No.11, Nov. 2004, pp.1480-1484 (2004)

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

[4] Y. Zaydun, Ts. Saitoh: "Sub-word Structure Analysis of Uyghur Language and Its Application to Character Recognition", IEICE Technical Report, Vol.105, No.500 (IE2005 127-185), pp. 47-50 (2006).

[5] Cha S.H., Y.C. Shin and S.N. Srihari: "Approximate Stroke Sequence String Matching Algorithm for Character Recognition and Analysis", in Proc. 5th Int. Conf. on Document Analysis & Recognition (ICDAR99), pp.53-56 (1999).

[6] M. Nakai, T. Sudo, H. Shimodaira and Sh. Sagayama: "Pen Pressure Features for Writer-Independent On-line Handwriting Recognition Based on Substroke HMM", in Proc. 16th Int. Conf. on Pattern Recognition (ICPR 2002), Vol.III, pp.220-223 (2002-08).

[7] Y. Zaydun, Ts. Saitoh: "The Application of Character Structure Information in Online Handwriting Uyghur Character Recognition", Journal of the IIEEJ 194, 2008, Vol.37, No.3, pp.244-249 (2008).

[8] 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).