A Syllable-based Associational Scheme for Online Handwriting Tibetan Character Recognition

Weilan Wang , Duojie Zhuoma, Bi Liang, Jianjun Qian

***Abstract*—An online handwriting Tibetan character recognition scheme of syllable-base association is proposed in the paper. First basic characteristics of the Tibetan language of letter, character, syllable and word are explained, and the different writing styles are introduced as well. At the some time, character's characteristics as recognition unit in our study are analyzed. In the proposed syllable association scheme, not only recognition speed but also recognition veracity is separately improved after isolated character recognized in our online handwriting Tibetan character recognition system. The research show, if adopt effectual method to improve the recognition rate of isolated character, the method will bring into play a more efficiency.**

Index Terms— **Online Tibetan character recognition, syllable-based associational, syllable structure, statistic, recognition rate**

I. Introduction

During recent years, the task of online handwriting recognition has gained an immense importance in every day application, mainly due to the increasing popularity of the personal digital assistant, electronic pen, electronic paper, or table PC notebooks [1][2]. And owing to the availability of both temporal stroke information and spatial shape information, online character recognition is able to yield higher accuracy than offline recognition, the target of recognition has shifted from regular script to fluent script in order to better meet the requirement of practical applications [3]. At the some time, with the widespread use of computer, the need for friendly human-machine interfaces is on rise [4], and different language man-machine interfaces. About Tibetan input in keyboard, we have developed an intelligent and past input system [5][6]. In the area of Tibetan character recognition, printed Tibetan character recognition has acquired some achievement of study [7][8] and practical application software has been used in Tibetan region. So far, however, online handwriting Tibetan character recognition still in the phrase of study, and there is very little work in this aspect. On the basis of previous work [9][10], in order to farther improve recognition performance, an associational scheme for online handwriting Tibetan character recognition is proposed in the paper, the method greatly improving efficiency of recognition.

The problem of online handwriting recognition is usually divided into four sub-problems: pre-processing (which can include segmentation), individual character recognition, candidate string generation, and post-processing. We focus post-processing part in this paper. This paper is organized as follows: in section 2, the characteristics of the Tibetan language script are introduced. Syllable-based associational Tibetan recognition scheme are discussed in section 3. Section 4 is experiment result and analysis. In section 5, we draw some conclusion and suggest direction for future work.

Manuscript received December 6, 2008. This work was supported in part by The Scientific Research Foundation for the Returned Overseas Chinese Scholars, Ministry of Personnel of People's Republic of China, and the Tackle Key Problems in Science and Technology Foundation of Gansu Province under Grant (NO: 2GS064-A52–035-04)

Weilan Wang was with School of Computer Science and Information Engineering, Northwest University for Nationalities, Lanzhou, Gansu Province,730030 China; e-mail: wangweilan@xbmu.edu.cn

Duojie Zhuoma was with Information Technology Institute, Northwest University for Nationalities, Lanzhou, Province,730030 China; e-mail: djzm868@sina.com.

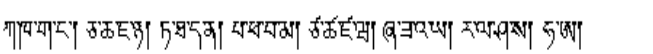
Bi Liang was with Information Technology Institute, Northwest University for Nationalities, Lanzhou, Province,730030 China. e-mail: liuhuaming888@126.com

Jianjun Qian., was with Information Technology Institute, Northwest University for Nationalities, Lanzhou, Province,730030 China. e-mail: q_jianjun830@yahoo.com.cn

II. Characteristics of the Tibetan Language Script

A. Letter, character, syllable and word in Tibetan language

There are 30 consonant letters:



And 4 vowels:  which are not separate written.

Syllabic structure in Tibetan language. A syllable may contain from one to seven letters. The simplest syllable consists of a isolated radical letter, while the most complex comprises seven letters in the following combination: a prefix (a prefixed letter placed immediately before the radical letter, 5 prefixes), a superscript (a superscripted letter placed immediately above the radical, 3 superscripts), a radical (corresponding to one of the 30 consonants), a subscript (a subscripted letter placed immediately below the radical letter, 4 subscripts), a vowel accent (a vowel accent placed above or below the radical consonant, there are 4 vowel accents, and the A is implied by default), a first suffix (a first suffix placed immediately after the radical) and secondary suffix (a second suffix placed immediately after the first suffixed consonant). Even if there is theoretically a very large set of possibilities,

the actual number of complex syllables is relatively limited. 5 prefixes are: ག,ད,བ,མ,འ ; 3 superscript letters are ར ལ ས, ར is distortion of ར ; 4 subscripts ཝ,ྱ,ྲ,ླ . ྱ,ྲ,ླ are distortion of ཡ,ལ,ར respectively; 10 first suffixes are ག,ང,ད,ན, བ , མ , འ , ར , ལ ས respectively; 2 secondary suffixes are: ད ,ས.

The principles of spelling. As a general rule, when a syllable contains several letters, they are spelled out in the following order: prefix, superscript, radical, subscript, vowel, first suffix, second suffix. In other words, the letters are spelled out horizontally from left to right and vertically from top to bottom (except in the case of superscripted vowels, which are pronounced after the subscripted consonant)

Syllable is basic language unit in Tibetan language contains single-character syllables, double-character syllables, triple-character syllables and quadruple-character syllables. For example: གྲོ ,བགྲོ ,བགྲོད, བགྲོད་པས which are single- character syllable, double-character syllable, triple-character syllable and quadruple-character syllable and just they have same root: ག.

There are 562 characters, which are vertically stacked by consonant letter and vowel, consonant letter each other and they and vowel themselves. This stacked layer number is not better than 4 in modern Tibetan. The character is a recognition unit in our study.

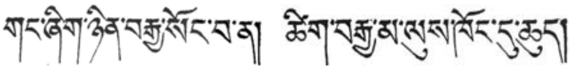
A word can be one syllable, two syllables or multi-syllables. A dieresis was used between different syllables. For example, the word སློབ་ཚན (lesson) combined by two syllables སློབ and ཚན.

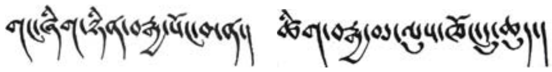
The sentence of Tibetan language can be formed by rules of syntax by words, but without blank between words. For instance: འདི་སྨྱུ་གུ་ནག་པོ་རེད། (means: this is a black pen.), the simple bar "།" express comma or a full stop.

B. The different styles of writing

The many styles of writing Tibetan may be grouped into two main categories: "with a head" or "white letters", used mainly for the purposes of printing, and "without a head" or "black letters", which includes the various cursive and ornamental styles. The cursive styles exceeding differ from printing styles. 5 lines in same sentences are showed in Fig, it indicate one type of the printing styles (order number: 1 and 2) and some cursive styles (order number: 3,4,5). Handwriting printing styles (order number: 1 and 2) are used in our study. Tibetan style is baseline alignment from left to right, the above vowels appear in the top position of baseline in a character.

(1) ཡང་ཞིག་ཉིན་བརྒྱ་སོང་བ་ན། ཚིག་བརྒྱ་མ་ལུས་ཁོང་དུ་ཆུད།

(2) 

(3) 

(4) 

(5) 

Fig 1. "with a head" and "without a head"

Fig 2 shows some handwriting samples as "with a head" or "white letters".

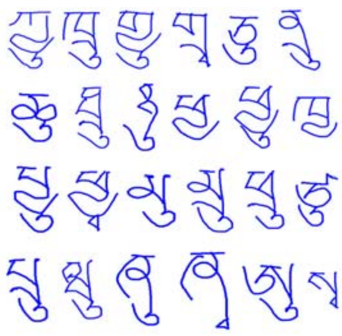


Fig 2. Tibetan writing samples

C. Character characteristics of Tibetan language

The main characteristics of character as follows:

(1) The amount of character is many. There are 562 characters in modern Tibetan language barring punctuation, symbol etc., and more than 3700 Sanskrit characters. We only study online handwriting the character of modern Tibetan language in this paper.

(2) The stroke is complex. There are many curve strokes and across strokes in character set. Fig 2 shows the character of simple stroke and the character of complex stroke.

(3) The highs of characters not equal. All consonants don't equal in high, such as consonant ལ and ཤ, character ཟ and སྨྱུ

(4) There are many similar pairs character in Tibetan language. Due to the similar of two above vowels, make about 100 similar character pairs, excess 1/3 of whole character set. For example སྲི and སྲེ, སྦྲི and སྦྲེ , etc. In addition, there are about 12 pairs very similarity characters due to the similar of consonant letter of composing character, namely similar pairs of བ and པ, other similar pair like ཟྲི and ཟྲེ, ཟ and ཐ, ཤྲི and ཥྲི etc.

(5) The writing habits

According to the writing habits of Tibetan language, writing begins from baseline position ordinal down. If a character take below vowel, usually last stroke is the below vowel; while if a character take above vowel, last stroke is the above vowel. Above vowels and below vowel can be written with one or two strokes. It some time is exceptional, however writing begins from above vowel, so the stroke order will be neglect in our study.

III. SYLLABLE-BASED ASSOCIATIONAL SCHEME FOR TIBETAN CHARACTER RECOGNITION

A. Associational rules

We have indicated in section of the 2.3 that there is a lot of similar character pairs in Tibetan language, so similar of character distinguish to be an important aspect to online handwriting recognition. For improve recognition rate and input speed, there are many ways can be used, one of means is to use language knowledge base on context-sensitive, or speaking, syllable-based associational function by Tibetan language rules in which mean syllable's spell. In our study, a syllable-based association recognition method is proposed. The rules of syllable-based association recognition are as following:

(1.) If a recognized character is one of 5 prefix letters, then possible syllables are: (i) "prefix + radical character + suffix + second suffix", such as འཕྲོགས; (ii) "prefix + radical character + suffix " such as འཕྲོག; (iii) "prefix + radical character" such as འཕྲོ;

(2.) If a recognized character is one of the 30 letters, possible syllable will be one type of following: (i) "radical" as a syllable such as ཀ; (ii) "radical + suffix" such as ཀག; (iii) "radical + suffix + second suffix" like ཀགས

(3.) If a recognized character is a stacked (involve only with vowel) character, possible syllable form: (i) single-character syllable like བྲི; (ii) double-character syllable that form is "radical character + suffix" such as བྲིས; (iii) triple-character syllable that form is "radical character + suffix+ second suffix" like བྲིགས.

Fig 3 shows the relationship between each element in a syllable that it is radical-center form by prefixed, suffixed, superscripted and subscript based on definite rules. Where, every component is marked with circle; straight or arc the arrowhead indicated the relation between each component in a syllable. Specially, thick arc denote the relation of prefix, radical, suffix, and secondary suffix; void arc denote the relation of prefix, radical and suffix; thin arc denote the relation of radical, suffix and secondary suffix. The three cases above (1),(2) and (3) can be explained by the figure. Where, characters take without vowel, only one above vowel or one below vowel.

B. Syllable-based associational recognition scheme

A syllable-based associational recognition block is shown in Fig 4.

The library file in Fig 4 can be created by associational algorithm in D section.

User can use associational method to input a syllable after character is recognized, by this time the " Library File" in action. If user only input single character one by one, then the input will be "Character Code" after "Character Classification ".

The associational method is much more efficient in our online handwriting Tibetan recognition. The most salient characteristic is fast input and the results will be show in Table 2 Ⅳ section.

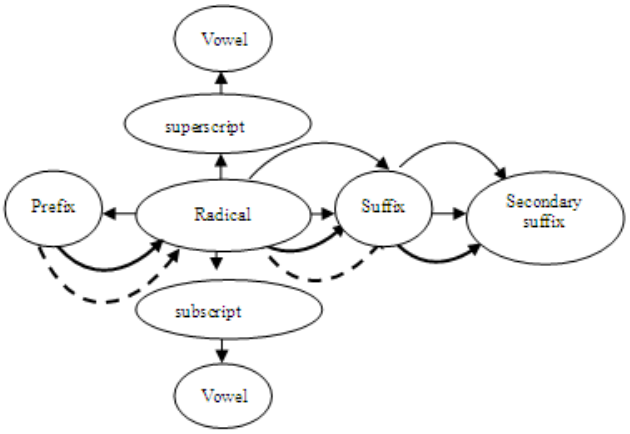


Fig 3. Relation between each element in syllable

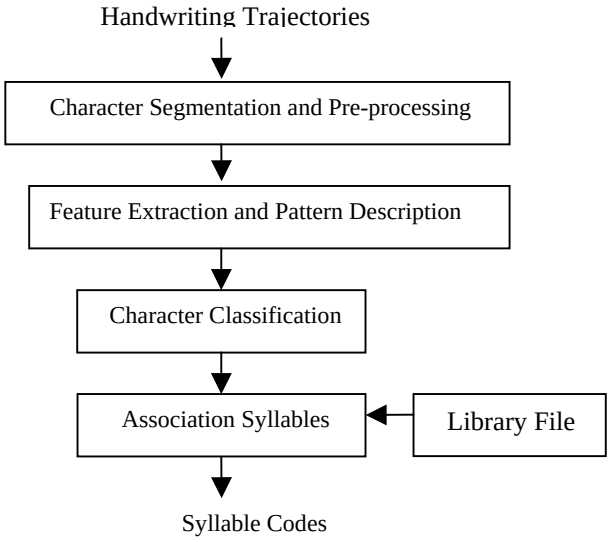


Fig 4. Block Diagram of Recognition System

Like recognition system of other character, our system of online handwriting Tibetan character recognition contains 5 parts, they are pre-processing, feature extraction, and character classification and post-processing respectively. We will be not discussed the front parts in this paper. Our emphases will focus on "Association Syllables". The output of system obtains 10 characters of candidate by reliability in pattern recognition section "Character Classification".

C. The statistic of syllable

The syllables, which accord with the spelling rules of syllable, are called theoretically syllables, and they can be generated by algorithm and program. We have finished the task; the total amount of syllable is 14892. Where, numbers of single-character syllables, double-character syllables, triple- character syllables and quadruple-character syllables are 445, 4985, 7212 and 2250, respectively.

The syllables that are obtained from statistical angle by corpus are called actual syllables which numbers are 415 single-character syllables, 2475 double-character syllables, 2423 triple-character syllables, 524 quadruple-character syllables respectively, and all together there are 5837. The numbers is far less the numbers of theoretically syllables. The statistical results of Tibetan syllables are in language databank with a size of 50 megabyte.

In a corpus that is containing 6709466 syllables, times, frequency and accumulative frequency of syllables of different length are studied respectively. As shown in Table 1, the syllable of the most frequently is "*pa*" in single-character syllable, 13.2% of all single-character syllables put together. Only 62 single-character syllables can cover about 90.2% and all the other 352 syllables account for less than 9% of the total amount of syllables, which are available in Tibetan language. Out of 2489 double-character syllables, only 305 syllables are responsible for the coverage of 90% of texts and the accumulative frequency of all the other 2184 double-character syllables account for less than 9% of the coverage of the texts in corpus. Triple-character and quadruple-character syllables also indicate non-uniformity in the frequency and accumulative frequency. At the same time, we concluded that only 12% of syllables in actual use appear in more than 90% of the texts.

Table 1. Frequency distributing of syllable

Order	Syllable	Times	Frequency%	Accumulative Frequency %
1	ཀ	258689	13.1814375	13.1814375
62	ཀྵ	5181	0.2639966	90.1533127
1	དུང	171606	6.5056510	6.5056510
305	ཅོར	1281	0.0485632	90.0341568
1	གཱིས	39529	3.1712427	3.1712427
251	བཟའ	753	0.0604100	90.0003433
1	དཔེགས	7991	9.9708023	9.9708023
61	འཕྲངས	203	0.2532941	90.0416718

Table 1, these statistical results indicate: (1) the number of the high frequency is less, 87.7% of the syllables that are covered within the corpus are normative and a great number of non-normative syllables exist; (2) Of more than fourteen thousand syllables, 39.2% is found to be syllables in actual use. And the theoretical value of these syllables of different length is respectively as following: single-character is 93.26%, double-character syllable 49.65%, triple-character syllable 33.6% and quadruple-character syllable 23.3%.

The statistical results show actual use of syllables is few than theoretic syllables and high frequency syllable much less, so after character recognized to utilize associational method to associate full syllable that are an excellent method.

D. Associational algorithm

A syllable-based associational algorithm as following:

Step1. Transform Tibetan character in the syllabic frequency list to the index number of character.

Step 2. All of syllables are divided into 562 character categories by the first character's index number, a Tibetan syllable belonged to the categories after syllable frequency list was ergodic dealt with.

Step 3. Count the numbers of syllable in each character categories, and sort all syllable in each character categories by frequency of syllable.

Step 4. Create an associational library file.

Step 5. Load the library file while startup the online handwriting Tibetan recognition system.

Step 6. Export all of Tibetan syllable with the character at its head by the index number of the recognized character in the character list.

IV. Experiments and Results

A. Specification of sample

300 sets samples, and 168600 character samples are collected from 300 Tibetan students' writing. 192 sets sample are used in the system.

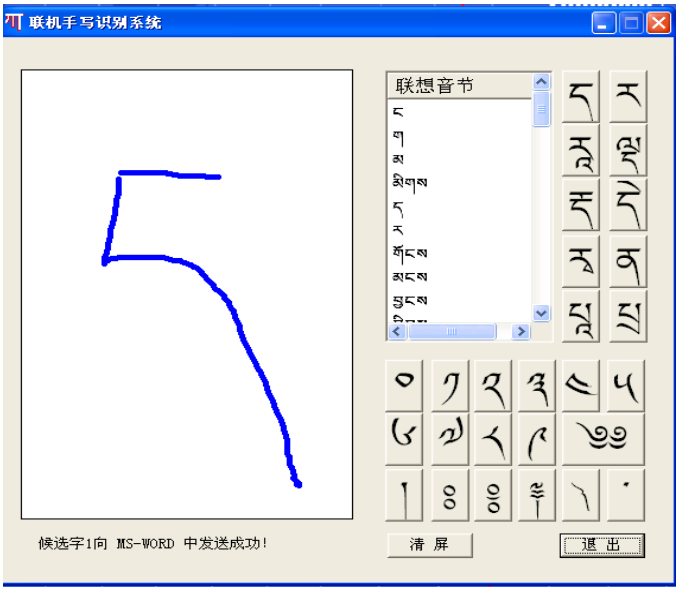
B. Results of isolated character

142 sets sample as training and 50 sets sample as test, the recognition rate of varied characters between 98% and 64% by test to isolated, and the average accuracy is 84.78% for top 10 candidates for test sample in our system.

C.Associational results

Fig 5 shows the interface of system. There are 4 parts: writing-box, results area of 10 candidates characters, association results area of syllable and pushbutton area of symbol containing numbers, special symbol and punctuation.

The first character of ད just is itself in candidates, associational syllables may are དང, དག, དན, དམིགས, དད and so on, all of syllables display in the scroll bar of rectangular



area, prefix is ད or radical is ད, by their frequency, described in A. (2) of Ⅲ.

Fig 5. The interface of recognition system

The associational method has very obvious effects in our online handwriting Tibetan recognition. For example, if we input character one by one not to select associational syllable, and use the results of associational syllable, will have a great time difference, the column of "Word" is some Tibetan words. The column of "Syllable number", "Character number" is numbers of syllable, numbers of character in

Table 2. Time difference of isolate character input and association of syllable input

Word	Syllable number	Character number	Time difference	Word	Syllable number	Character number	Time difference
མཚོངས	1	4	10	ཡིན་ན་ཡང	3	7	13
དགོངས	1	4	8	སྨན་ལམས་རེ	3	8	4
དྲངས	1	3	7	སྲུབས་མེད་ཐོག	3	9	9
སྐོག་གྲུབ	2	5	4	ཁྱད་པར་འཕགས་པ	4	12	17
སྐྱབས་དཔྱིད	2	8	13	རྒྱ་ནག་སྐྱས་རི	4	10	6
སྐད་གད་ངས	2	7	15	གེམས་རུན་ཐམས་ཅད	4	13	13
སྐུ་ལམས	2	5	10	རྒྱལ་བའི་ཐུགས་སྐས	4	12	6
ལྷ་ཆེན་པརྒྱད	3	8	15	རྒྱ་མའི་རྒྱང་ཚངས	4	11	9

a word respectively. The column of "Time difference" T denotes the time difference of both T_1 and T_2:

$$T = T_1 - T_2$$

Where, T_1 : consuming time when input the word by character recognition one by one .

T_2 : consuming time after character recognition to use associational syllables.

V. CONCLUSIONS

In this paper, a syllable-based associational online handwriting Tibetan language character recognition scheme is proposed, the method improve the performance both recognition rate and recognition speed. Its efficiency will be very high while isolated character recognition rate attain a high level. Therefore our aim is farther improvement recognition rate of isolated character to adopt different methods such as HMM, the combined method of online and offline handwriting recognition and so on, we will farther consider online handwriting recognition of syllable as well.

REFERENCES

[1] Thierry Artieres, Sanparith Marukatat, and Patrick Gallinari. Online Handwriting Shape Recognition Using Segmental Hidden Markov Models. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, FEBRUARY 2007. VOL. 29, NO. 2.pp 299-310.

[2] Claus Bahlmann and Hans Burkhardt. The Writer Independent Online Handwriting Recognition System frog on hand and Cluster Generative Statistical ynamic Time Warping. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, MARCH 2004. VOL. 26, NO. 3.pp 299-310.

[3] Cheng-Lin Liu, Stafan Jaeger and Masaki Nakagawa. Online Recognition of Chinese Characters: The State-of-the-Art. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, FEBRUARY 2004. VOL. 26, NO. 2.pp 198-213

[4] Karteek Alahari, Satya Lahari Putrevu, and C.V. Jawahar. Discriminant Substrokes for Online Handwriting Recognition. Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition (ICD AR'05).

[5] WANG Wei-lan. Intelligent Input Software of Tibetan. Computer Standards & Interfaces. 29 (2007) pp462–466

[6] Weilan Wang, Lingwang Kun. A Fast Input Method for Tibetan Based on Word in Unicode. The International MultiConference of Engineers and Computer Scientists 2008.pp375-377

[7] Ding Xiaoqing, Wang Hua. Multi-Font Printed Tibetan Character Recognition. JOURNAL OF CHINESE INFORMATION PROCESSING. Year:2003 Issue:06 Volume: 17

[8] Wang Weilan Ding Xiaoqing Chen Li, Wang Hua. Study on Printed Tibetn Character Recognition. Computer Engineering. 2003. Vol. 29. pp37-38,94

[9] Liu Hongyi, Wang Weilan. Nonlinear Shape Normalization Methods for On-line Recognition of Handwritten Tibetan Characters. Application Research of Computers. 2006(9).pp179-181.

[10] Wang Weilan, Chen Wan-jun. MCLRNN Model for Online Handwritten Tibetan Character Recognition Based on Stroke Characteristics. Computer Engineering and Applications. 2008.Vol.26, No.14.pp 91-91,194