Research on the Method of Tibetan Recognition Based on Component Location Information

Yuehui Han[1,2], Weilan Wang[1(✉)], Yiqun Wang[1], and Xiaojuan Wang[1]

[1] Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lanzhou 730000, Gansu, China
wangweilan@xbmu.edu.cn

[2] College of Mathematics and Computer Science, Northwest Minzu University, Lanzhou 730000, Gansu, China

Abstract. The recognition of Tibetan is of great significance to the study of Tibetan culture while the progress of Tibetan character recognition is lagging behind. Especially when there are not a large number of available training samples, Tibetan character recognition is very difficult. So we propose a recognition method for Tibetan characters based on component location information without a large number of training samples. The proposed method includes three main parts: (1) The segmentation of character and the extraction of component which contain location information in the character; (2) Features extraction and classifier design; (3) The superposition of component after recognition and the retrieval of character. The testing results are: the recognition rate of single component is 98.4%, the recognition rate of multilevel component is 97.2%. It indicates that the method has a good effect on the recognition of Tibetan character, and it is helpful for the recognition of Tibetan documents.

Keywords: Tibetan recognition · Character segment
Component combination · Classifier design

1 Introduction

Tibetan is a minority nationality character which is used by 5 million Tibetan people in China. There are two views on the origin of Tibetan character: One view is that the Tibetan was created by a minister Tumi Sabza of Srongtsen Gampo's in the seventh Century. Another view is that the Tibetan was evolved from Zhang zhung character. Tibetan is a special kind of phonetic character, whose longitudinal unit is a character, and a character consists of at most 4 components. Syllables are the basic spelling units. Each syllable consists of at most 4 characters, as shown in Fig. 1.

Compared with other languages, the progress of Tibetan recognition research is relatively backward. However, the gap is gradually narrowing under the efforts of a lot of scholars.

J.-H. Lai et al. (Eds.): PRCV 2018, LNCS 11258, pp. 63–73, 2018.
https://doi.org/10.1007/978-3-030-03338-5_6

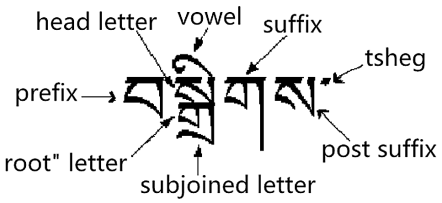


Fig. 1. Example of Tibetan structure

In Printed Tibetan: Hua Wang carried on the preliminary study of Tibetan recognition form the preprocessing, text line segmentation, feature selection and classifier design [1]. By using the segmentation method based on the connected domain and the extraction of the stroke feature based on the grid, Zhu Ou increased the recognition rate of the Tibetan [2]. In order to improve the recognition rate, Yulei Wang extracted the features of Tibetan characters based on Fractal Moments and improved rough mesh method [3]. Yuzhen Baima proposed projection method based on network lattice which is suitable for Tibetan recognition [4]. Wei Zhou proposed a Tibetan recognition method based on geometry analysis of component [5]. In Handwritten Tibetan: Heming Huang established the first off-line handwritten Tibetan recognition system [6]. Xiaojuan Cai proposed a feature extraction algorithm for off-line handwritten Tibetan characters based on multi projection normalization, which further improved the recognition rate [7]. By using HMM based on stroke type and the position relation between strokes to improve the recognition performance [8], Weilan Wang designed a complete online handwritten Tibetan recognition system [9], proposed a Tibetan Sanskrit handwritten sample generation method based on component combination [10]. Longlong Ma proposed a semi-automatic component annotation method for online handwritten Tibetan character database [11], a Tibetan component representation learning method for component-based online handwritten Tibetan character recognition [12], and a component segmentation-based recognition method for online handwritten Tibetan syllables [13]. We propose a recognition method for Tibetan characters based on component location information without a large number of training samples. The rest of this paper is organized as follows.

Section 2 introduces printed Tibetan characters and components. Section 3 gives the component segmentation method. The method of feature extraction and classifier design is given in Sect. 4. Section 5 gives recognition process and result analysis. Section 6 offers concluding remarks.

2 Tibetan Characters and Tibetan Components

Tibetan is a special kind of alphabetic writing that a character contains 1 to 4 components which are superposed up and down. Most Tibetan recognition work is based on characters, while the recognition work based on components is rarely. There are 534 printed Tibetan characters used frequently, while 231 components in totally. And the 231 component contains 51 single components, 180 deformation combination

components. As for non-single that changes have taken place in the deformation combination, so we take combination components as a whole, as shown in Table 1 and Fig. 2. In fact based on components is a very useful method for Tibetan recognition work especially when the training sample is insufficient. Tibetan characters have strict distribution rules, which can help separation component easily. Based on component can also help reduce the number of classification. Character is recognized by retrieving Tibetan characters database after the components are recognized.

Table 1. Example of Tibetan characters database.

ID	Tibet	TibetOrder	Sort	Code
144		82	1	41
145		82	2	3
146		82	3	161

Table 1 is a character example in Tibetan characters database, "TibetOrder" is the sequence number of the character in database, "Tibet" is a character, "ID" is the database record number, "Sort" is the layer information of a component in a character, and "Code" is the sequence number of component in the template. Figure 2 is all Tibetan components which contain 51 single components and 180 deformation combination components.

We proposed a recognition method for Tibetan characters based on components location information. The stages of the proposed method are shown as follow.

(1) After the size transformation, the segmentation of the above vowel, the segmentation of the below vowel and the segmentation of intermediate component, the component containing location information are obtained.
(2) Feature extraction and classifier design.
(3) Calculate the matching degree using the Euclidean distance, screen out the top-ten matching degree and the corresponding components.
(4) According to the recognition result of each component, retrieve and find out the corresponding character in database.

3 Component Segmentation

Component segmentation based on the writing standard of Tibetan character, which follow the sequence of above vowel, below vowel and intermediate component. The component segmentation process is shown in Fig. 1.

In Fig. 3, "Above" indicates above vowel, "Below" represent below vowel and "Single" indicates single intermediate component, "Double" refers to double intermediate component.

Fig. 2. All Tibetan components we used.

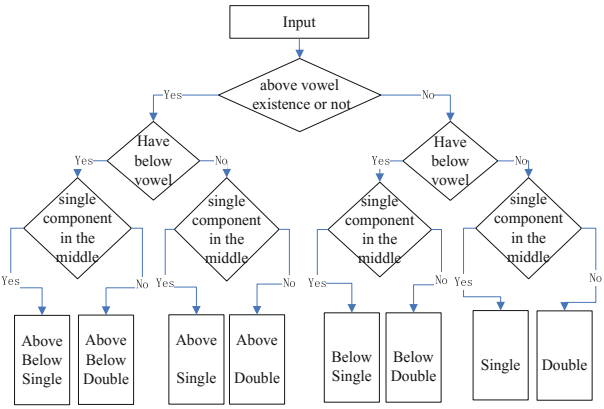


Fig. 3. Component segmentation process

3.1 Above Vowel Segmentation

The above vowel is located on the baseline of the Tibetan character, the top 1/4 part of an image. The above of the baseline is empty without above vowel. The specific algorithms are as follows.

Step 1: Above vowel judgment.

The statistical number of handwriting points in the 1/5 section above the image, and the numbers is replaced by "sup". Column projection on the 1/5 section above the image, Statistical the numbers that Greater than zero, and the numbers is replaced by "tnum". The method of judgment is shown in Fig. 4, "Cnum" represent the numbers of columns. Experimental verification, when T is 5, there is the best result.

Step 2: Find the segmentation point.

Image line projection, and the point near "Rnum/4", which has minimum projection value and has the maximum rate is the segmentation point. "Rowsnum" indicate the numbers of lines.

Step 3: Above vowel segmentation.

Image segmentation based on segmentation point. Example of above vowel segmentation is shown in Fig. 5.

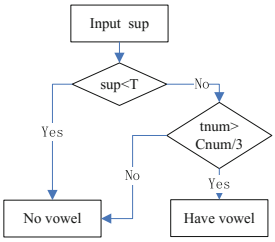


Fig. 4. Above vowel judgment

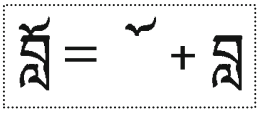


Fig. 5. Example of above vowel segmentation

3.2 Below Vowel Segmentation

The below vowel is located in the underneath, 1/4 part of image. The specific algorithms are as follows.

Step 1: Below vowel judgment.

The statistical number of handwriting points in the bottom 1/5 section of the image is replaced by *sdown*. Column projection on the 1/5 section bottom the image, Statistical the numbers that Greater than zero, and the numbers is replaced by "dnum". The method of judgment is shown in Fig. 6.

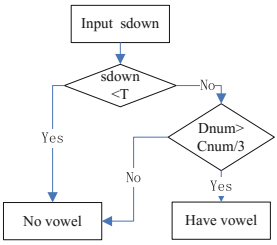


Fig. 6. Below vowel judgment

The numbers of "Cnum" indicate the number of columns.

Step 2: Find the segmentation point.

Projection in the right half of the image, and the point near "4*Rnum/5", the segmentation point is supposed to have minimum projection value. "Rnum" indicate the numbers of lines.

Step 3: Below vowel segmentation.

Starting from the right side of the image, if connected to a below vowel, disconnect based on the segmentation point. If not connected, search the segmentation path along the contour of below vowel. Figure 7 is the Example of below vowel segmentation.

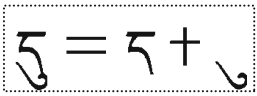


Fig. 7. Example of below vowel segmentation

Intermediate component Segmentation

There are only one or two layers of components in middle part, after above vowel segmentation and below vowel segmentation. The specific algorithms are as follows.

Step 1: Judgment of the number of layers.

After removing the above vowel and below vowel, assume the number of handwriting points in the top half of the image is N, in the bottom half of the image is M. Single component if $M/N < T_3$, the middle part is called single component, and it is called double component under the condition of $M/N > T_3$. The experiment proves that the result is best when M is 0.9.

Step 2: Find the segmentation point.

Projection the image, and the point near the middle position of the image, which has minimum projection value is the segmentation point.

Step 3: Intermediate component segmentation.

Image segmentation is based on segmentation point. Example of intermediate component segmentation is shown in Fig. 8.

Fig. 8. Intermediate component segmentation

3.3 Special Circumstances Process

(1) Sometimes the segmentation of above vowels may makes mistakes, as is shown in Fig. 9. In this case we can use the minimum rectangle to extract the correct top component. As is shown in Fig. 10.

(2) Sometimes the deformation combination of some components will be considered as a single component, which is shown in Fig. 11. So we consider the result of the deformed combination as a component and increase the number of components in the template.

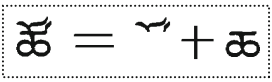


Fig. 9. Error segmentation example

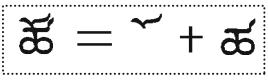


Fig. 10. Correct segmentation example

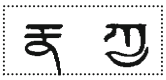
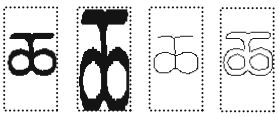


Fig. 11. Component deformation combination

4 Feature Extraction and Classifier Design

4.1 Component Feature Extraction

168 features are extracted altogether, and the images involved are original component image, remove position information image, skeleton image and edge image. As is shown in Fig. 12(a)–(d). All image normalization, 100 rows and 50 columns.



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

Fig. 12. (a) (b) (c) (d) Image used to extract feature

The feature extraction algorithm of the component is as follows.

Step 1: Feature extraction of original image

The original image refers to the component image come form template or character segmentation. The original image contains the location information of the component distribution. And the distribution information of different components is different. Four features are extracted from the original image: The ratio of black pixel points, the number of rows with black pixel points, the position of first and the last row with black pixel points.

Step 2: Feature extraction of remove position information image

After minimum rectangle frame processing, image extends to the original size. And the image is divided into 16 parts using an elastic grid. 23 features are extracted from the remove position information image: The ratio of black pixel points, position of grid line and the position of first black pixel point per line in each part.

Step 3: Feature extraction of skeleton image

After skeleton processing of the original image, we get the skeleton image. 41 features are extracted from the skeleton image: rough periphery and inner profile.

Step 4: Feature extraction of edge image

After edge processing of the original image, we get the edge image. And the image is divided into 25 parts averagely. Statistical directional line information in each part and 100 features are extracted.

4.2 Classifier Design

Euclidean distance is used to calculate the matching degree between the test components and the components in the template. D_i indicate the matching degree between the test components and the i-th components in the template. And the range of number "i" is 1 to 231. As shown in (1).

$$D_i = \sum_{j=1}^{m} (x_j - x_{i,j})^2 \tag{1}$$

Where m indicate the total number of feature values, x_j and $x_{i,\ j}$ represents the j-th feature value of test component and the j-th feature value of i-th components in the template.

5 Analysis of Experimental Results

The method is carried out after line and character segmentation. Figure 13 is a part of the Tibetan document image. Figure 14 is line segmentation results. Figure 15 is the recognition results of Fig. 14(a), and the results are "གདན་ས་བཀྲ་ཤིས་ལྷུན་པོ།". It can be seen from the recognition example that our method has a satisfactory recognition result. For the experiments 100 Tibetan printed document images are used, and the recognition rate of single component is 98.4%, the recognition rate of multilevel component is 97.2%.

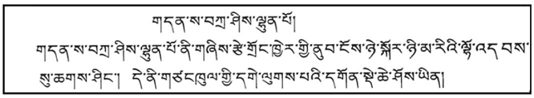
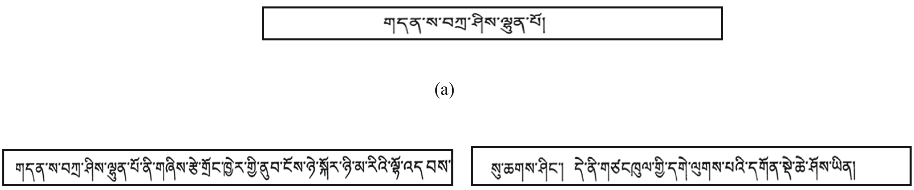

གདན་ས་བཀྲ་ཤིས་ལྷུན་པོ།
གདན་ས་བཀྲ་ཤིས་ལྷུན་པོ་ནི་གཞིས་རྩེ་གྲོང་ཁྱེར་གྱི་ནུབ་ངོས་ཉེ་སྐོར་ཉི་མ་རིའི་སྟོ་འདབས་
སུ་ཆགས་ཤིང་། དེ་ནི་གཙང་ཁུལ་གྱི་དགེ་ལུགས་པའི་དགོན་སྡེ་ཆེ་ཤོས་ཡིན།

Fig. 13. Tibetan printed document example


གདན་ས་བཀྲ་ཤིས་ལྷུན་པོ།

(a)

གདན་ས་བཀྲ་ཤིས་ལྷུན་པོ་ནི་གཞིས་རྩེ་གྲོང་ཁྱེར་གྱི་ནུབ་ངོས་ཉེ་སྐོར་ཉི་མ་རིའི་སྟོ་འདབས་

སུ་ཆགས་ཤིང་། དེ་ནི་གཙང་ཁུལ་གྱི་དགེ་ལུགས་པའི་དགོན་སྡེ་ཆེ་ཤོས་ཡིན།

(b) (c)

Fig. 14. (a) (b) (c) line segmentation results

Character	Component segmentation result	Recognition results Top1 ->Top10
ག	ག	ག ས ན ནྟ ཀ ཏྲ ཤ ཞ ཟ
ང	ང	ང ཀྵ ཏ ཀ ར ཀྱ ག ཡ ར ན
ན	ན	ན ནྟ ཅ ན ས ཏ ཀྵ ས ག ཅ
་	་	་ ༢ ཚོ ཀ ཇ ཀྵ ཀྵ ཚ ཚ ཀོ
ས	ས	ས ད ག ལ ས ཉ ཚ ད ཤ ཌ
་	་	་ ༢ ཇ ཀ ཚོ ཀྵ ཚ ཁྲ ཚ ཀྵ
ད	ད	ད ཌ ཐ ད ཤ ཐ ཏ ཐ ས ཌ
ཀྱ	ཀྱ	ཀྱ ཏ ཏ ཏ ཁ ད ཏ ཤ ཁྲ ཐ
་	་	་ ཚོ ༢ ཀ ཇ ཀྵ ཀྵ ཚ ཚ ཀོ
ཅི	ི ཀ	ཅི ཆི ཀྵི ཅེ ཅི ཅོ ཅཾ ཅཱ ཅོ
ས	ས	ས ཉ ད ག ས ཚ ད ཉ ཁ ཏ
་	་	་ ༢ ཚོ ཇ ཀ ཀྵ ཚ ཀྵ ཚ ཁྲ
ཀྵེ	ེ ྱ ྀ	ཀྵེ ཀྵ ཁ ཀྵ ཀྵ ཀྵ ཀྵ ཀྵ ཀྵ ཀྵ
ན	ན	ན ནྟ ན ཅ ས ག ཀྵ ད ས ས
་	་	་ ཚོ ༢ ཀ ཇ ཀྵ ཀྵ ཚ ཚ ཁྲ
ཏོ	ོ ཏ	ཏོ ཏཱ ཏི ཏེ ཏོ ཏཾ ཏཱ ཏི ཏེ ཏོ
ག	ག	ག ནྟ ན ཅ ས ག ཀྵ ད ས ས
་	་	་ ཚོ ༢ ཀ ཇ ཀྵ ཀྵ ཚ ཚ ཁྲ
ཀྱོ	ོ ཀྱ ཨ	ཀྱོ ཀྱོ ཀྱ ཀྱོ ཀྱོ ཀྱོ ཀྱོ ཀྱོ ཀྱོ ཀྱོ
ན	ན	ན ནྟ ན ཅ ས ག ཀྵ ད ས ས
་	་	་ ཚོ ༢ ཀ ཇ ཀྵ ཀྵ ཚ ཚ ཁྲ
ཏོ	ོ ཏ	ཏོ ཏཱ ཏི ཏེ ཏོ ཏཾ ཏཱ ཏི ཏེ ཏོ
།	།	། ་ ན ཀ ཅ ཏ ག ཤ ཏ ཁ

Fig. 15. Recognition example

After analysis, we can see that there are two main reasons for wrong recognition. One is caused by the line or character segmentation error, which may cause some character information to lose or increasing noise. And another is caused by components segmentation error, which segmentation points are judged mistakenly. Wrong segmentation point cause wrong segmentation results and lead to wrong recognition results certainly, which is the reason why multi-layer character recognition rate is lower than single-layer character.

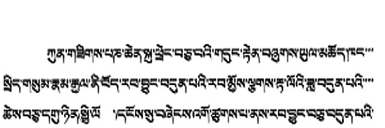


Fig. 16. Black body

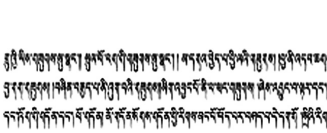


Fig. 17. Long body

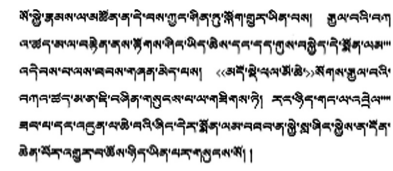


Fig. 18. Round body

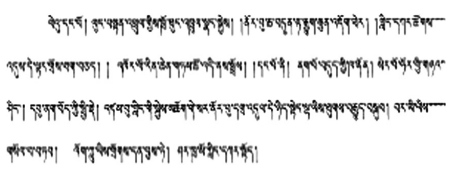


Fig. 19. Bamboo body

We also tested four other Tibetan fonts that include Black body (see Fig. 16), Long body (see Fig. 17), Round body (see Fig. 18) and Bamboo body (see Fig. 19), which recognition rate is 96.3%, 92.1%, 95.8% and 93.3% in the 50 sets of test samples. From the test results, we can see that the recognition effect of Black body and Round body is better than Long body and bamboo body. This is because that the change of Long body and Bamboo body is larger than that of Black body and Round body compared with the commonly used Tibetan fonts, which is the template we use. So it is easy to make mistakes when components are segmented, which lead to the component contain noise or some information lost. And then the result of the character recognition is wrong. Although these Tibetan fonts are slightly different from the commonly used Tibetan fonts, but the recognition rate has not been greatly affected. This also can prove that the characteristics extracted are effective.

6 Conclusions

This paper propose a recognition method for Tibetan characters based on component location information when lack a large number of training samples. The main work includes: the extraction of component which contain location information, features extraction based on four kinds of images, classifier training, superposition of component and the retrieval of character based on component location information database. The single-layer character recognition rate for this method is 98.4%, and 97.2% for multi-layer character. It is found that the effect of component segmentation directly affects the recognition of character. So the optimization of component segmentation algorithm is the focus of further research.

Acknowledgements. This work was supported by the National Science Foundation (No. 61772430), the Program for Leading Talent of State Ethnic Affairs Commission, the Fundamental Research Funds for the Central University of Northwest Minzu University (No. 31920170142), and also supported by the Gansu Provincial first-class discipline program of Northwest Minzu University.

References

1. Wang, H., Ding, X.Q.: Multi-font printing Tibetan character recognition. J. Chin. Inf. Process. **17**(6), 47–52 (2003)
2. Drup, N., Ren, P., Sanglangjie, D.: Study on printed Tibetan character recognition. Comput. Eng. Appl. **48**(1), 55–62 (2009)
3. Li, Y.Z., Wang, Y.L., Liu, Z.Z.: Study on printed Tibetan character recognition technology. J. Nanjing Univ. **48**(1), 55–62 (2012)
4. Baima, Y.Z.: Research on feature extraction of Tibetan characters. Comput. Knowl. Technol. **28**(1), 6362–6364 (2013)
5. Zhou, W., Chen, L.: Tibetan recognition based on geometric shape analysis. Comput. Eng. Appl. **48**(18), 201–205 (2012)
6. Huang, H.M.: Research on recognition of off-line handwritten Tibetan character, pp. 19–34. Southeast University (2014)
7. Cai, X.J., Huang, H.M.: Feature extraction of offline handwritten Tibetan characters based on multiple projections. Comput. Technol. Dev. **26**(3), 93–96 (2016)
8. Liang, B., Wang, W.L., Qian, J.J.: Application of hidden Markov model in on-line recognition of handwritten Tibetan characters. Microelectron. Comput. **26**(4), 98–100 (2009)
9. Research on online handwritten Tibetan recognition input: W.L. Wang. Sci. Technol. Achiev. China **11**, 36–38 (2012)
10. Wang, W.L., Lu, X.B., Cai, Z.Q.: Online handwritten sample generated based on component combination for Tibetan-Sanskrit. J. Chin. Inf. Process. **31**(5), 64–73 (2017)
11. Ma, L.L., Wu, L.: Semi-automatic Tibetan component annotation from online handwritten Tibetan character database by optimizing segmentation hypotheses. In: 12th International Conference on Document Analysis and Recognition, pp. 1340–1344 (2013)
12. Ma, L.L., Wu J.: A Tibetan component representation learning method for online handwritten Tibetan character recognition. In: 14th International Conference on Frontiers in Handwriting Recognition, pp. 317–322 (2014)
13. Ma, L.L., Wu, J.: Online handwritten Tibetan syllable recognition based on component segmentation method. In: 13th International Conference on Document Analysis and Recognition. pp. 46–50 (2015)