A Component Recognition-based Chinese Character Segmentation and Structure Discrimination Method

Yongtang Bao*, Yue Qi, Bowen Yu
State Key Laboratory of Virtual Reality Technology and Systems
School of Computer Science and Engineering, Beihang University
Beijing, China
Author*: baozi0221@163.com

***Abstract*—In this digital age, using the technology of computer graphics and image processing to analyzing Chinese character segmentation and structure discrimination, has a very important significance for the inheritance and development of traditional culture. In this paper, we present a Chinese character segmentation and structure discrimination method based on component recognition. First, we use contour detection and boundary tracing method to trace component contour. We segment the input character component by finding the correspondence between contour points and skeleton points. Second, we analyze the common Chinese character structure and the golden grid construction theory. Finally, we propose the structure discrimination standards. Our method can produce accurate Chinese character segmentation and structure discrimination results.**

Keywords—component recognition; Chinese character segmentation; structure discrimination; golden grid construction theory

I. INTRODUCTION

The problem of Chinese character recognition is involved by pattern recognition and image processing discipline, which is the technique with strong comprehensiveness. It has both values on applications and theory meanings in the field of Chinese information processing, automatically working and machine intelligence. So many researches on statistical method, structural method, statistical along with structural approach are proposed in recent years, which make the technique of Chinese character recognition develop fast and the effect of Chinese character recognition further improved.

Wang et al. [1] processed the Chinese character image as a whole. They recognized the Chinese character based on feature vector value representing the information of whole image. Pal et al. [2] recognized the character by the distance in statistical method. Araki et al. [3] proposed a statistical approach for character recognition using Bayesian filter. This method represented the character by creating Hash table, and calculated the Bayesian joint probability to recognize character. These statistical approaches can easily extract feature, but cannot perfectly recognize the Chinese character in similar shape.

The first step for structural character recognition method is the extraction of Chinese character strokes. The stroke extraction contains two types, top-down and bottom-up strategies [4-5]. The structural method can reflect the structural characteristics, and can recognize the character easily. But this method cannot robust, and it is difficult to extract structural element from image.

The statistical structure method synthesize the advantage of statistical method and structural method, it is robust and can adapt deformation. Yu et al. [6] proposed an improved statistical structure modeling method to pick all meaning components in one character. Each stroke is represented by the distribution of the feature points both in model component and input character. Sun et al. [7] constructed a teaching system using statistical structure theory. This method reduces the requirement for element extraction, and it can achieve better recognition effect.

With the results of Chinese character component recognition, we segment the Chinese character and discriminate the structure of it. The organizational structure of this paper is: (1) We introduce contour detection and boundary tracing method in section II. We segment the Chinese character component based on correspondence among contour points and skeleton points. (2) In section III, we analyze the common Chinese character structure and the golden grid construction theory. (3) We give the structure discrimination standards in section IV. The results are shown in Section V.

II. CHINESE CHARACTER SEGMENTATION

A. Contour Detection

Contour is composed of linking edge points, containing abundant image information. Contour is the main way of representing image feature. In this paper, we use Canny edge detection operator for contour detection. The Canny edge detection contains three steps: smoothing, enhancement and detection.

In smoothing step, we use Gauss filter method to smooth original image. Equation (1) and (2) gives the 1D and 2D Gaussian kernel function. The kernel functions will be used to convolute original image after normalization.

$$K_1 = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \quad (1)$$

$$K_2 = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (2)$$

After enhancement and detection, the result may contains too much noise, we need to use double threshold and edge connection to complete full edge detection.

B. Edge Tracing

The result of contour detection method is without connection information for preceding and subsequent edge points. We use the edge tracing method to find the contour points. We solve this problem according to eight connected region.

1	0 1	7 0	7
2	A 2	6 B	6
3	4 3	5 4	5

Fig. 1. Target point and its eight neighborhood points.

As shown in Fig. 1, the point right above the target point is 0 neighborhood point. The eight neighborhood points are marked from 0 to 7 in anticlockwise manner. If A is the X neighborhood point of B, then B is the (X + 4) neighborhood point of A (in octal form). For example, A is the 2 neighborhood point of B, and B is the 6 neighborhood point of A.

C. Contour and Skeleton Point Correspondence

After getting the contour points and skeleton point, we need to find the correspondence among them. After computing the correspondence, we extract the contour points and connect them at the disconnected position to generate the final segmentation results.

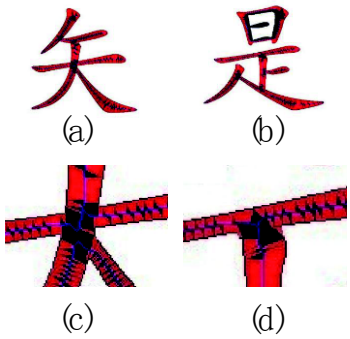

Fig. 2. Map for contour and skeleton point correspondence

As shown in Fig. 2, (c) and (d) are the results of cross section of contour and skeleton points for Chinese character "矢" and "是".

III. GOLDEN GRID CONSTRUCTION THEORY

A. Common Chinese Character Structure

As shown in Fig. 3, the Chinese character can be classified into thirteen common structure. For example, the Chinese character "思" is composed of two components, "田" and "心". The location relationship of these two components is up-down, so the Chinese character "思" is up-down structure. The structure of all common Chinese character can be found in Fig. 3.

Num	Pattern	Structural Name	Examples
1		up-down structure	思、杏
2		up-middle-down structure	怠、草
3		left-right structure	休、明
4		left-middle-right structure	弼、班
5		all surrounding structure	囚、困
6		left-up-right surrounding structure	同、冈
7		left-down-right surrounding structure	凶、函
8		left-up surrounding structure	病、原
9		up-left-down surrounding structure	匠、区
10		left-down surrounding structure	毯、建
11		right-up surrounding structure	司、句
12		independent entity structure	大、由
13		mosaic structure	噩、巫

Fig. 3. Thirteen common Chinese character structure

B. Golden Grid Construction Theory

A Chinese character is written in a target box, the box is divided into nine boxes based on golden section method. Fig. 4 shows the target box after golden section.

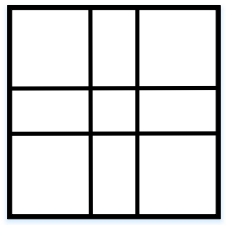

Fig. 4. Golden grid of Chinese character

IV. STRUCTURE DISCRIMINATION STANDARDS

A. Component Combination and Structure Discrimination

After character segmentation in section II, we can extract a group of components. We analyze the Chinese structure based on the golden grid construction theory. For each component, we judge which of the nine box are occupied by it and give the discrimination standard for each structure.

The result of character component segmentation is various subcomponents, while the destination of structure discrimination is for getting entirety by combining the subcomponents recursive. During the combining, we analyze the structure relationship of these two subcomponents, then we can get the hierarchical structure of Chinese character constitute.

The flow chat of Chinese character component combination and structure discrimination can be seen in Fig. 5. We select the most appropriate two component to combine, according to position, length and width similarity of bounding boxes.

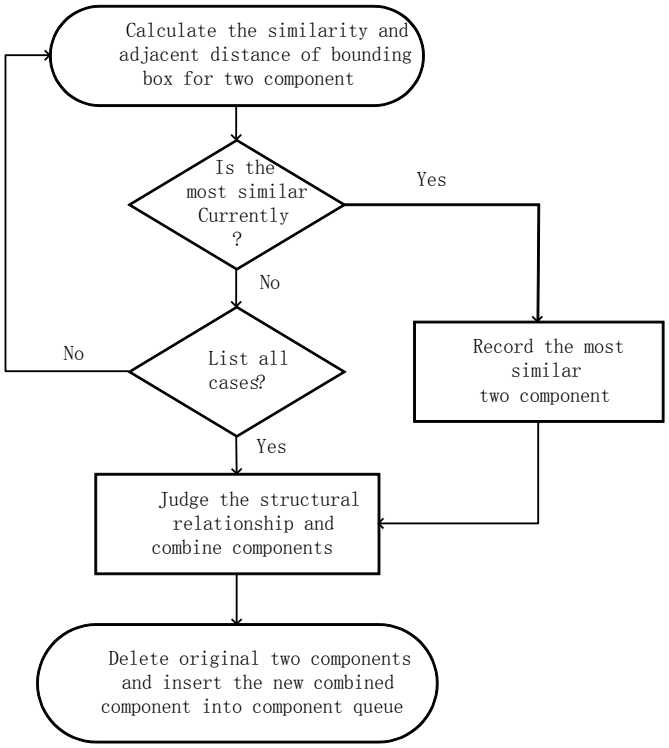

Fig. 5. Flow chart of component combination and structure descrimination

We describe the procedure of judging the structural relationship of two combining components as following: (1) Combine these two components and calculate the bounding box of it; (2) Extend this whole sub-image to 400*400 size and place it to the central of 500*500 image; (3) Divide the 500*500 image with golden grid and calculate the occupancy rate of these two components by nine grid; (4) Calculate the occupancy rate of these two components by the central grid in golden grid, and set the component major occupied the central grid as centrobaric component; (5) Complete the structure discrimination by the occupancy rate of another component. During judging, we set the grid value to 1 if one component is occupied by this grid and set it to 0 if the component is not occupied by this grid. (6) Multiply the grid value with the grid weight (can be seen in Fig. 6) to get the index value for structure discrimination. The structure discrimination standards are described in next section.

1	8	32
2	0	64
4	16	128

Fig. 6. The grid weight of golden section box

B. Structure discrimination standards

For each thirteen common Chinese character, we define the discrimination standard as following:

1) Up-down structure: For each component, we calculate the index value as sum of products for the occupied golden grid and corresponding weight. If one component has occupied the centrobaric grid, the index value of another component can be calculated as 1+8+32=41 or 8+32=40, if the index value of this component is 41, 40, 9, 144, 208, 22, 104, 11, 148, 43, 105, 107, 150, 214 or 212, we judge these two component as up-down structure.

2) Up-middle-down structure: For three components, we judge the location relationship two by two, if they are all the up-down structure, the structure for these three component is up-middle-down.

3) Left-right structure: Similar to up-down structure, if the index value of the component is 7, 224, 15, 23, 31, 232, 240, or 248, the two components are left-right structure.

4) Left-middle-right structure: Similar to up-middle-down structure, if the last twice judging result are all left-right structure, these components are left-middle-right structure.

5) All surrounding structure: The index value is 255.

6) Left-up-right structure: The index value is 111, 235 or 239.

7) Up-left-down structure: The index value is 63, 159 or 191.

8) Left-down-right structure: The index value is 215, 246 or 247.

9) Left-up surrounding structure: The index value is 47.

10) Right-up surrounding structure: The index value is 233.

11) Left-down surrounding structure: The index value is 151.

12) Independent entity structure: The input Chinese character is composed of only one component, such as “大”.

13) Mosaic structure: If it cannot be judeged as above twelve cases.

Using these thirteen discrimination standards, we can judge the hierarchical structure relationship for any input Chinese character.

V. RESULTS

A. Component Recognition

Fig. 7 shows the results of component recognition for some Chinese character. The results will be used for component segmentation and structure discrimination.

Fig. 7. Results of component recognization

B. Component Segmentation

As shown in Fig. 8, we give the component segmentation results for some Chinese characters. Fig. 8 (a) is the ordinary segmentation. Fig. 8 (b) and (c) are need to segment at the cross point. Fig. 8 (d) demonstrates the multiple components segmentation.

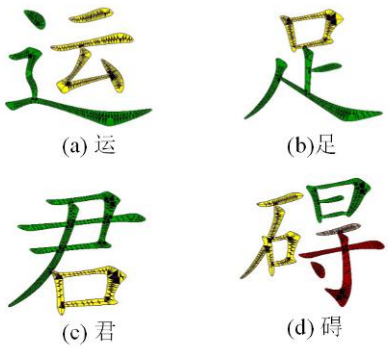

Fig. 8. Results of component segmentation

C. Structure Discrimination

Fig. 9 shows the results of structure discrimination. The Chinese character "囚" belongs to the all surrounding structure according to the discrimination standards in previous section. The Chinese character "匹" is left-up-down structure. "茬" and "喳" are composed of multiple components. The hierarchical structure of them can also be seen from Fig. 9.

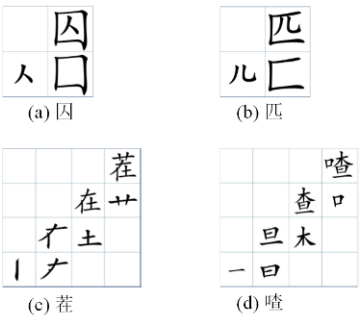

Fig. 9. Results of structure discrimination

VI. CONCLUSION

We have present a Chinese character segmentation and structure discrimination method based on component recognition. We use the golden construction theory to segment character. We also propose the structure discrimination standards for analyzing character structure. Demonstrated with some Chinese characters, our method could generate accurate results.

ACKNOWLEDGMENT

This paper is supported by National Key Technology Research & Development Program of China (2014BAK18B01).

REFERENCES

[1] Y. Wang, X. Ding, C. Liu, "MQDF discriminative learning based offline handwritten Chinese character recognition," Document Analysis and Recognition (ICDAR), 2011 International Conference on. IEEE, 2011, pp.1100-1104.

[2] S. Pal, J. Mitra, S. Ghose, "A Projection Based Statistical Approach for Handwritten Character Recognition," Proceedings of the International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007)-Volume 02. IEEE Computer Society, 2007, pp. 404-408.

[3] N. Araki, M. Okuzaki, Y. Konishi, "A statistical approach for handwritten character recognition using bayesian filter," Innovative Computing Information and Control, 2008. ICICIC'08. 3rd International Conference on. IEEE, 2008, pp. 194-194.

[4] J. Kim, H. Kim, "Statistical character structure modeling and its application to handwritten Chinese character recognition," Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2003, 25(11), pp. 1422-1436.

[5] M. Su, F. Wang, "Decomposing Chinese characters into stroke segments using SOGD filters and orientation normalization," Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on. IEEE, 2004, pp.351-354.

[6] B. Yu, X. Liang, J. Hu, "Statistical Structure Modeling and Optimal Combined Strategy Based Chinese Components Recognition," Signal Image Technology and Internet Based Systems (SITIS), 2012 Eighth International Conference on. IEEE, 2012, pp.238-245.

[7] L. Sun, M. Liu, J. Hu, "A Chinese Character Teaching System Using Structure Theory and Morphing Technology," PloS one, 2014, 9(6): e100987.