

## Segmentation of Handwritten Chinese Character Strings Based on improved Algorithm Liu

Zhihua Cai, Jiangqing Wang\*, Yangguang Sun

College of Computer Science, South-Central University for Nationalities, Wuhan 430074, China

\*Corresponding author, e-mail: hunterpaper2199@gmail.com

### Abstract

*Algorithm Liu attracts high attention because of its high accuracy in segmentation of Japanese postal address. But the disadvantages, such as complexity and difficult implementation of algorithm, etc. have an adverse effect on its popularization and application. In this paper, the author applies the principles of algorithm Liu to handwritten Chinese character segmentation according to the characteristics of the handwritten Chinese characters, based on deeply study on algorithm Liu. In the same time, the author put forward the judgment criterion of Segmentation block classification and adhering mode of the handwritten Chinese characters. In the process of segmentation, text images are seen as the sequence made up of Connected Components (CCs), while the connected components are made up of several horizontal itinerary set of black pixels in image. The author determines whether these parts will be merged into segmentation through analyzing connected components. And then the author does image segmentation through adhering mode based on the analysis of outline edges. Finally cut the text images into character segmentation. Experimental results show that the improved Algorithm Liu obtains high segmentation accuracy and produces a satisfactory segmentation result.*

**Keywords:** Character Strings segmentation, Character recognition, Connected components analysis, Merged characters segmentation

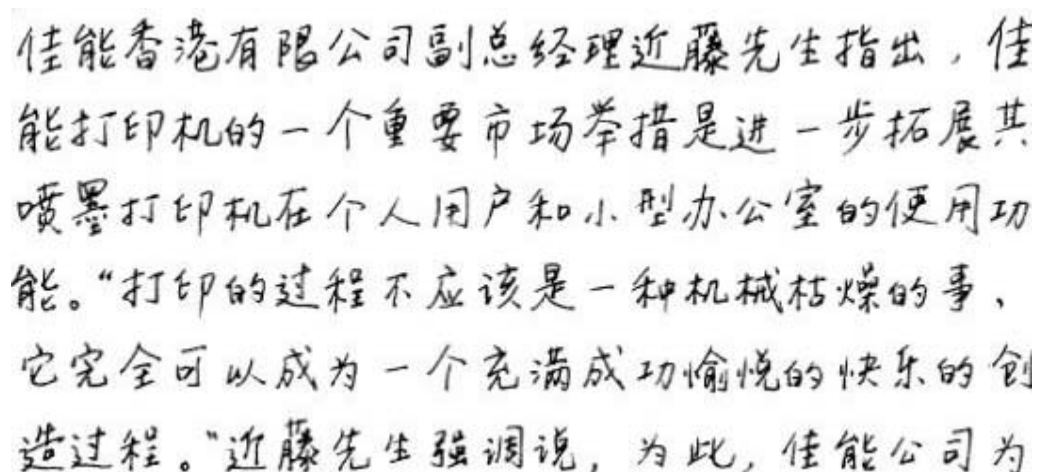
### 1. Introduction

Character segmentation is one of the key technologies in text recognition rate. The object of character segmentation can be divided into printed character, handwritten Chinese character, handwritten western character, etc. It is need to choose different segmentation methods for different object generally. The scholars and experts at home and abroad do a lot of researches on segmentation of western character and numbers [1] and have made a lot of valuable achievements. But Chinese character segmentation always focuses on printed character. There are few researches on handwritten Chinese character segmentation, so the corresponding achievements are also few. Because Handwritten Chinese character is different from printed Chinese character. Briefly, printed character is mechanical segmentation of two dimensional planes, which is standard and easy to recognize. While handwriting is lively, vivid and writing model with high artistry, and handwriting pays attention to the shape, generalization and change of stroke image. What's more, there must be natural and fluent coherent and echo between stroke and stroke. Therefore, there are conditions of script in handwriting Chinese character and strokes are changing, as shown in Fig.1. There is certain tilt in the font in this handwriting image and different spacing between line and line. Meanwhile, the space between word and word is different. Some are written relatively tensely, while others are written relatively sparse. These characteristics are all distinctive from the printed text. Therefore, the segmentation of handwritten Chinese character is more complicated.

Certainly, there are some achievements which are worth noticing. A segmentation method which is based on stroke segment drawing is proposed by Tseng L.Y., etc. [2]. They adopts regulatory computing cost matrix of components merger based on knowledge, and then obtain text image segmentation by using dynamic programming. Chen Hong et al. [3] make use of the characteristics of width of characters and spacing, etc. to do character segmentation in text through the principle of minimum variance. Zhao Yuming et al. [4] propose a handwritten Chinese character segmentation method based on stroke extraction and merging. Zhao et al. [5-6] propose two-stage segmentation method of coarse segmentation and fine segmentation. Jiangqing Wang et al. [7] propose vertical projection characters segmentation based on minimum threshold and curve-fitting. Zhao Shuyan etc. [8] propose merged handwritten

Chinese character segmentation based on stroke analysis and background thinning. Cao Wei [9] proposes segmentation of gap algorithm based on multi-threshold and multi-segmentation strategy. Li Xiaoyuan etc. [10] propose merged handwritten Chinese character segmentation based on structural cluster analysis and stroke analysis. These methods promote the recognition of Chinese character to a certain extent, but there are still some shortcomings, such as recognition accuracy and recognition speed, etc., because recognition of Chinese character is under the influence of some objective factors, such as writing quality, writing style, font size, and image quality of handwritten Chinese character on computer, etc. Algorithm Liu [11] is proposed by Cheng-Lin Liu, Masashi Koga and Hiromichi Fujisawa. It is acquired with high accuracy in Japanese postal address segmentation, but its algorithm is complex and its implementation is difficulty, which have negative effects on its popularization and application.

In this paper, the author applies the principles of algorithm Liu to handwritten Chinese character segmentation according to the characteristics of the handwritten Chinese characters, based on deeply study on algorithm Liu. In the same time, the author put forward the judgment criterion of Segmentation block classification and adhering mode of the handwritten Chinese characters. In the segmentation process, the text line is considered as the sequence of CCs and each CCs is comprised of a set of black runs. It is determined whether CCs is merged or not by analysis of CCs. Then, the segmentation block are segmented into character segmentation block by touching patterns segmentation based on analysis of partial contours. Experimental results show that the new algorithm has a high precision rate.



佳能香港有限公司副总经理近藤先生指出，佳能打印机的一个重要市场举措是进一步拓展其喷墨打印机在个人用户和小型办公室的使用功能。“打印的过程不应该是一种机械枯燥的事，它完全可以成为一个充满成功愉悦的快乐的创造过程。”近藤先生强调说，为此，佳能公司为

Figure 1. The image of Handwritten Chinese character

## 2. Algorithm Liu

Liu algorithm is the method of character segmentation based on connected components [11]. In the process of segmentation, text images are seen as the sequence made up of Connected Components (CCs), while the connected components are made up of several horizontal itinerary set of black pixels in image. And then analyze each connected component recursion. The analysis process is that calculate the standard overlap. If the standard overlap of the two parts is positive, this two parts can be merged. Calculating the standard overlap is similar to calculating relevancy to the two parts. Only the standard overlap is high enough, and it thinks that the two parts can be merged. The components which complete the merger are called text segmentation. Because text segmentation after merging exist adhesion owing to handwriting, that is, segmentation after merging contains two character, conducting adhering mode segmentation based on outline edges analysis. Segmentation process is to do outline analysis of text segmentation which is inconsistent with height-width ratio, and detect point of division. If the point of division is not found, judge again whether text segmentation conforms to height-width ratio of compulsory segmentation. If it meets, do compulsory segmentation of it. If not, give up compulsory segmentation. If you could find the point of division, do segmentation

on this text segmentation. Don't do segmentation on those which are consistent with height-width ratio. Finally, get the complete text segmentation. Algorithm Liu is effective for segmentation of handwriting text image which exist adhering mode, and find point of division and complete segmentation accurately.

This algorithm is a complete recognition system of postal address, including image preprocessing, connected component analysis, adhering mode segmentation and a search algorithm based on the classifier of character and the dictionary. The system flowchart is shown in Figure 2:

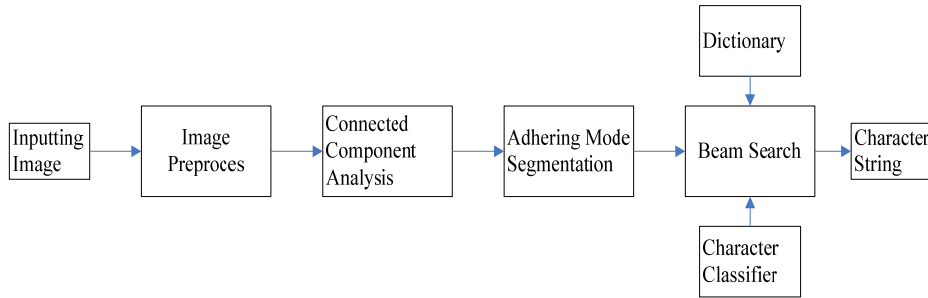


Figure 2. The Character String Recognition (CSR) System

The whole system begins with a text image. Its writing direction is from left to right or from top to bottom. The text image is presented by Connected Components (CCs). The system will process text image in sequence until it can output an address phrases. The accuracy of these 3589 actual email texted by the research group is almost to the point of 83.86% and the error rate below 1%. Visibly, this algorithm is effective for the segmentation of Japanese postal address.

### 3. The Segmentation of Handwritten Chinese Character Based on Algorithm Liu

#### 3.1. Image Preprocessing

There may be some factors, such as fuzziness and alteration, etc. affecting handwritten text, and the problems, such as un-clarity and noise, etc. in its image. Therefore, to estimate stroke width of text image from Run Length Histogram of Connected Components in order to control image filtering effect. Firstly, to find stroke width-- $rl_{max}$ , which appears the most, that is, the maximum probability from Run Length Histogram. But this value does not represent the stroke width of handwriting text image, because there may be other noise or other information contributing to this probability. Therefore, the mean value of studying run length histogram,  $rl_{mean}$ , is calculated by the formula 1 below:

$$rl_{mean} = \text{int} \left[ \frac{\sum_i i \cdot hist_1(i)}{\sum_i hist_1(i)} \right] \quad (1)$$

Among them,  $\text{int}(\cdot)$  is round off the mean value calculated. If  $rl_{max} \geq rl_{mean}$ ,  $SW = rl_{max}$ , otherwise, find a value which satisfies the formula 2 between  $rl_{max}$  and  $rl_{mean}$ .

$$\begin{aligned} hist_1(rl_{local}) &\geq \frac{1}{2} hist_1(rl_{max}) \\ SW &= rl_{local}, \text{ or } SW = rl_{max}. \end{aligned} \quad (2)$$

Estimating stroke width of text through runs length histogram. All CCs units which are smaller than  $SW \times SW$  in the black area are removed. If  $SW < 3$ , will not do any further smoothing process on text image in order to prevent the fracture of thin lines; Otherwise, if  $SW \geq 7$ , the image will do morphological open handle by using the structure of  $5 \times 5$ ; otherwise, if  $SW \geq 5$ , the image will do morphological open handle by using the structure of  $3 \times 3$ ; if it is no less, the image will do median filter smoothing in the area of  $3 \times 3$ . Doing morphological open handle is effective for eliminating filament and noise. The structure template used is shown in Figure 3:

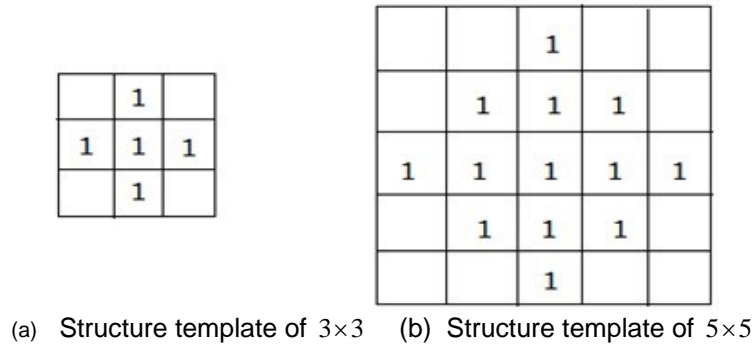


Figure 3. Structure template used by morphology

Stroke width ( $SW$ ) is the key to get binary image. Once it cannot get correct value of  $SW$ , it will affect the final result after entire algorithm processing. To do segmentation of character in image, the wrong stroke width will delete or dispose of the black pixels contained in the image, and then binary image obtained cannot reflect the text messages correctly, and of course, it cannot complete text segmentation.

### 3.2. Connected Components Analysis

Handwritten text image is seen as a sequence of CCs, while each CCs is composed of horizontal stroke set of several black pixels. Because CCs overlap is likely to constitute the same character, calculate its standardized overlapping degree to decide whether to merge it. Calculation process is shown below. Bounding box of components is appointed by coordinate of left, right, upper and lower boundaries. The bounding box of the two parts are respectively  $(x_1^l, x_1^r, y_1^l, y_1^b)$  and  $(x_2^l, x_2^r, y_2^l, y_2^b)$ .

Suppose  $x_1^l \leq x_2^l$ , if  $x_2^l < x_1^r$ , the two parts overlap, the overlapping degree is  $ovlp = x_1^r - x_2^l$ , span is  $span = \max(x_1^r, x_2^r) - \frac{dist}{span}$ ,  $nmovlp$  is calculated by formula 3 below.

$$nmovlp = \frac{1}{2} \left( \frac{ovlp}{w_1} + \frac{ovlp}{w_2} \right) - \frac{dist}{span} \quad (3)$$

Among them,  $w_1$  and  $w_2$  respectively represent the width of two parts,  $dist$  represents the horizontal distance of the center. If  $nmovlp$ , merge the two parts. As long as there is large enough standardized overlapping degree, to merge the current parts and its subsequent parts. Figure 4 shows diagrammatic sketch of merging two parts.  $O_1$  and  $O_2$  represent the center of the two parts. Merger of components adopts recursive combined method. After recursive merger, each part can be seen as image segmentation. And then detect the potential adhering mode of segmentation and cut it. If you think segmentation is still likely to be multi-model after segmentation is completed, you can force to cut the block according to its projection histogram.

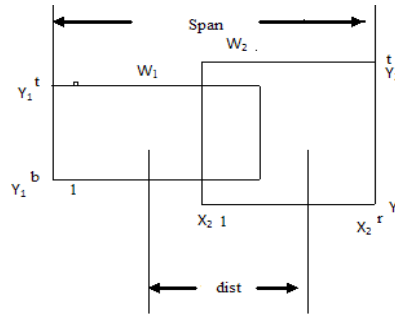


Figure 4. Arc strokes consisting of two segments

Connected components are the basic unit for analyzing the text segmentation. There are script and broken in writing. According to the definition of the connected components in the algorithm, connected components are identified as the set of a series of consecutive black pixel point. All points which can be connected can form a connected component in the text image. Here connected points are not only the points connecting up, down, left and right, but also the points connecting the upper left, lower left, upper right and lower right, that is, eight connecting area. Because characters are not written under the situation of “pure” horizontally and “pure” vertically, but written with a certain angle. Determination of connected components is the foundation of merging connected components. By adopting the method in this section, it could merge connected components into segmentation block, and then do adhering mode analysis next.

### 3.3. Adhering Mode Segmentation

1) The text segmentation block concluded from by the previous two steps does not represent the final result of segmentation. Therefore, there may be “under-segmentation” in text segmentation block at this time. So-called under-segmentation is that the two segmentation blocks which should be independent merge because of the too close writing. This merger is not the defect which algorithm itself design, but the writing. So, detect the adhering mode of segmentation and find out the point of division, in order to separate the two segmentation blocks which are under-segmentation.

First estimate the height of text line ( $LH$ ) according to histogram of image segmentation block after CCs merging. The segmentation block produced after connected components analysis combining can be divided into big segmentation block and small segmentation block. The standard of judging small segmentation block is determined by the maximum frequency of the height of segmentation block calculated. The segmentation block which is smaller than  $1/2$  is regarded as small segmentation block. The reason is that Chinese character is relatively homogeneous. When the segmentation blocks have cut characters out through combination of the connected components analysis, their height will be more uniform. There are several punctuation marks whose height is small after being split out. But the number of segmentation block of punctuation marks is relatively less than that of segmentation block of text. Therefore, it could determine directly the height of small segmentation block according to the height of the segmentation block with the largest frequency. Given the height of small segmentation block had a little effect on character height, just merging segmentation block temporarily, and then the height and width of temporary segmentation are respectively  $th$  and  $tw$ . The histogram  $hist_2(th) = hist_1(th) + tw$  is updated. The height of text line is regarded as an estimated value of the height mean value of temporary block in the histogram. Suppose the width of segmentation block is  $SW$ , the height is  $SH$ . If  $SW > \theta_{h1} \cdot LH$  or  $SW / SH > \theta_{h2}$ , this segmentation block is likely to be adhering mode.

The height-to-width of Chinese characters is generally stable because Chinese characters are relatively homogeneous. Because there are few characters which are too wide and with height within the text line height  $LH$ , or it is almost impossible to exist, it is likely to merge multiple connected components into a text segmentation block for adhesion in the

connected component analysis. It can determine the value of parameters,  $\theta_{h1}$  and  $\theta_{h2}$ , through a lot of experiments. It is reasonable to think that there is adhesion in the text segmentation blocks whose height-width ratios don't conform to these two formulas. Use the domain-specific knowledge for segmentation of adhering mode. The author of algorithm Liu extends the adhering mode proposed by H.Ikeda, etc. [12]. He puts forward seven types of horizontal writing, and six types of vertical writing, as shown in Figure 5.

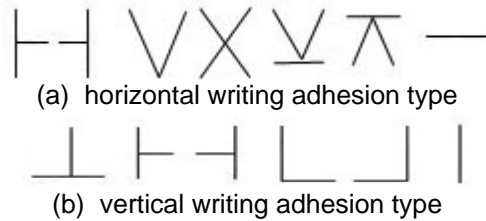


Figure 5. Two types of adhesion

2) Detect division point through analysis of local contour shape. The adhering mode is divided into single stroke area and many stroke more area. And then do contour analysis of each single stroke area. Using the technology proposed by Rosenfeld and Johnston [13] to detect the corner point for partial contour.

3) After the local contour shape analysis and segmentation, determine whether each segmentation block need forced segmentation. Forced segmentation is produced because Chinese character is a kind of relatively homogeneous Chinese characters. Therefore, the segmentation block through the second step of connected components analysis and merger is not likely to conform to height-width ratio, that is, the width is too wide, which is almost impossible to exist in the Chinese characters (as shown in Figure 6).

Write multiple width of handwritten line segmentation  $mw$  as the span from far left to far right in many stroke areas. If  $mw > \theta_{h3} \cdot LH$ , this segmentation block need forced segmentation. Define Eigen-Function the formula 4 below:

$$f(x) = \left[ \sum_y \bar{b}(x, y-1)b(x, y) \right] \sum_y b(x, y) \quad (4)$$

Among them,  $b(x, y)$  represents the binary image of segmentation block, and take the position where  $f(x) + |x - x_c|$  ( $x_c$  is the center of the multiple width) gets the minimum as the cut-off point. After completing pre-slitting, have one or more successive segmentation blocks combined as candidate character mode. Tag the space between the characters before dictionary matching.



Figure 6. Diagrammatic Sketch of Height-width Ratio of Text Segmentation Block

### 3.4. Analysis of the Segmentation

In this section, analyzing the process of the Segmentation by one sample, which is shown in Figure 7:

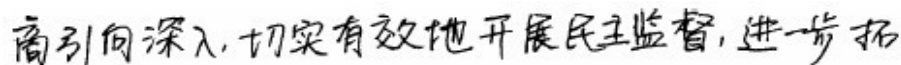


Figure 7. Experimental Sample Images

It can see from the sample image in Figure 7 that this sample image consists of those words “商引向深入, 切实有效地开展民主监督, 进一步拓”, which contain script, individual character and punctuation mark, almost covering several testing situations of handwritten Chinese characters.

It will analyze the sample image by stages below.

1) The experimental results of image preprocessing, as shown in Figure 8

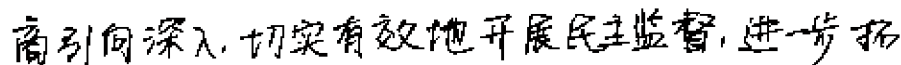


Figure 8. the Experimental Results of Image Preprocessing

After image preprocessing, text line image is composed of each black pixel point, which is different from the text line image with different gray value. In the experiment, it gets the stroke width  $SW$ , which is the two pixel point. Therefore, it doesn't process, and convert the original image into binary image.

2) The experimental results of connected components analysis, as shown in Figure 9:



Figure 9. the Experimental Results after Connected Components Analysis

In analyzing connected components, initially merge connected components recursive into text segmentation block one by one through calculating the standardization overlapping degree  $nmovlp$ . In Figure 9, the characters with left and right organization are written with relatively apart, so that those can't be merged into one text segmentation block. Thus make them be identified as a word in dictionary matching. Those text segmentation blocks with adhering mode, such as “民主” cannot be divided into two segmentation blocks and be regarded as a text segmentation block, because the final cross of “主” is linked to “民”. Therefore, divide text segmentation with adhering mode into two segmentation blocks or compulsorily to find the division point through adhering mode segmentation. For example, the two words “一步” also have the same problem. The segmentation block of punctuation marks obtained from above is very small, which contributes little to height value of the whole text line. It finds 35 connected components and 23 image segmentation blocks in the experiment. It gets the initial segmentation block of text line image through connected components analysis.

### 4. The Experimental Results and Analysis

Because test a large number of sample images in the experiment, only showing the part of results of the segmentation below, the sample images used are shown in Figure 10, Figure 11 and Figure 12.



品经营企业建立了“索证索票”等进货检查验证:

Figure 10. Experimental sample image

本报哈尔滨1月20日电 春节前夕,哈尔滨市

Figure 11. Experimental sample image

据了解,目前,北京市对食品安全已实现从田间到

Figure 12. Experimental sample images

The words in sample images respectively are “品经营企业建立了‘索证索票’等进货检查验证”, “本报哈尔滨1月20日电 春节前夕, 哈尔滨市”, “据了解, 目前, 北京市对食品安全已实现从田间到”. In sample images, these images contain the common situations, such as all kinds of font-style, script and punctuation marks, etc. in handwriting. The corresponding experimental results are shown in Figure 13, Figure 14 and Figure 15:

品经营企业建立了“索证索票”等进货检查验证

Figure 13. Images after Segmentation

本报哈尔滨1月20日电 春节前夕, 哈尔滨市

Figure 14. Images after Segmentation

据了解, 目前, 北京市对食品安全已实现从田间到

Figure 15. Images after Segmentation

It can conclude from the above experiments that algorithms Liu is very effective for segmentation of handwritten text line image.

## 5. Conclusion

The experiments show that algorithms Liu avoids the influence of noise and the information having nothing to do with text segmentation and ensures the segmentation quality. Based on the judgment criterion of Segmentation block classification and adhering mode according to the characteristics of the handwritten Chinese characters. The algorithm is more effective and accurate for text segmentation with adhering mode. The seven types in horizontal and six types in vertical ensure the search of division point of adhering mode. What's more, the parameters,  $\theta_{h1}$ ,  $\theta_{h2}$  and  $\theta_{h3}$ , in the experiment can be obtained from much experimental



statistics. This adapts to the characteristics of samples, is more targeted, ensures the segmentation effect and makes the algorithm have a better applied cost.

### Acknowledgments

This work is supported by the National Natural Science Foundation of China (60672057, 60975021), the Natural Science Foundation of South-Central University for Nationalities (YZY10006), the Open Foundation of Key Laboratory of Education Ministry of China for Image Processing and Intelligence Control (200906), the Science and Technique Research Guidance Programs of Hubei Provincial Department of Education of China (B20110802), the Fundamental Research Funds for the Central Universities", South-Central University for Nationalities (CZY12007, CTZ12013), and the Open Foundation of State Key Laboratory of Software Engineering (SKLSE20120934), the Natural Science Foundation of Hubei Province of China (2012FFB07404), the graduate Academic Innovation Foundation of South-Central University for Nationalities. The authors would also like to thank the anonymous reviewers for their valuable comments and constructive suggestions, which helped improve the quality of this paper.

### References

- [1] Xiaoyuan LI, Gang Tian, Chao Feng. Segmentation of Touching Characters in Printed Mathematical Expression. *Science Technology and Engineering (In Chinese)*. 2011; 11(3): 628-632.
- [2] Tseng LY, Chen RC. Segmentation Handwritten Chinese Characters Based on Heuristic Merging of Stroke Bounding Boxes and Dynamic Programming. *IEEE Pattern Recognition Letters*. 2007; 19: 963-973.
- [3] Chen Hong. Segmentation and Recognition of Continuous Handwriting Chinese Text. *International Journal of Pattern Recognition and Artificial Intelligence*. 1998; 12(2): 223-232.
- [4] Zhao Yuming, Jaing Xingzhi, Shi Pengfei. Algorithm for off-line handwritten Chinese character segmentation based on extracting and knowledge-based merging of stroke bounding boxes. *Infrared and Laser Engineering(In Chinese)*. 2002; 31(1): 23-27.
- [5] SY Zhao, ZR Chi, PF Shi, H Yan. Two-stage Segmentation of Unconstrained Handwritten Chinese Characters. *Pattern Recognition*. 2003; 36: 145-156.
- [6] SY Zhao, ZR Chi, PF Shi, Q Wang. Handwritten Chinese Character Segmentation Using a Twotage Approach. *Document Analysis and Recognition*. 2001; 6: 179-183.
- [7] Jiangqing Wang, Wei Cao. Vertical Projection Characters Segmentation Based on Minimum Threshold and Curve-Fitting. *Journal of South-Central University for Nationalities (Natural Science Edition)*. 2011; 30(4): 82-85.
- [8] Zhao Shuyan, Guo Jie, Shi Pengfei. Segmentation of Connected Handwritten Chinese Characters Based on Stroke Analysis and Background Thinning. *Journal of Shanghai Jiaotong University (In Chinese)*. 2003; 37(9): 1434-1437.
- [9] Cao Wei. Clearance Segmentation Based on Multi-threshold and Multi-segmentation Strategy. *Computer & Digital Engineering (In Chinese)*. 2011; 39(1): 131-133.
- [10] Li Xiaoyuan, Yang Fang, Zhang Wangbo. Segmentation of connected handwritten Chinese characters based on structure cluster and strokes analysis. *Computer Engineering and Applications (In Chinese)*. 2008; 44(34): 163-165.
- [11] Cheng-Lin Liu, Masashi Koga, Hiromichi Fujisawa. *Lexicon Driven Segmentation and Recognition of Handwritten Character Strings for Japanese Address Reading*. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2002; 11: 1425-1437.
- [12] Ikeda, Hisashi. *A Recognition Method for Touching Japanese Handwritten Characters*. International Conference of Document Analysis and Recognition. 1999; 5: 641-644.
- [13] A Rosenfeld and E Johnston. Angle Detection on Digital Curves. *IEEE Trans.Computers*. 1973; 22: 875-878.