International Journal of Multimedia and Ubiquitous Engineering
Vol. 9, No. 12 (2014), pp. 129-136
http://dx.doi.org/10.14257/ijmue.2014.9.12.12

Research of Chinese Handwritten Text Segmentation Algorithm

Zou Yingyong[1,2] , Zhang Yongde[1], Cao Xinyan[3] and Yu Guangbin[1]

[1]*Intelligent Machine Institute, Harbin University of Science and Technology, Harbin 150080, China*
[2]*Mechanical Engineering College, Changchun University, Changchun 130022, China*
[3]*School of Electronic and Information Engineering, ChangChun University, Chang Chun, China*
yyzou@126.com, zhangyd@hrbust.edu.cn

Abstract

OCR is a complicated process, there are many factors that can influence the recognition rate. Early period people tried to optimize the classifier to obtain high recognition rate, but the premise is that there is only one character no matter print or handwritten. For the performance of classifier has been promoted a lot, recognition rate for single character is high enough for commercial use. With the development of the demand for handwritten text recognition, how to raise the recognition rate of OCR system becomes very important. Unlike OCR system for print which focus on classifier. The research of OCR system for handwritten text is mainly on character segmentation. Statistical analysis showed that the mistake made by missegment is more than the mistake made by classifier. This is decided by the feature of handwritten text. There are more randomness and the lines are not horizontal, besides that, handwritten Chinese characters are more like overlapped and the gaps between characters are smaller. So this is the difficulty of handwritten Chinese characters. In this paper, the mutil-step searching nonlinear line exaction algorithm the paper proposed is easy and the accuracy is high, which can tackle the some weaknesses of direct projection method and indirect projection.

Keywords: *Chinese handwritten text, image segmentation, image recognizing, genetic algorithm*

1. Introduction

Chinese character is a tool for communication, and paper medium are traditional medium for recording Chinese characters. With the expansion of recognition technology, texts images need to be processed are more and more complex. Although people's desire for processing handwritten text grows with each passing day. After many studies, people found that mistakes caused by inappropriate segmentation are more than these caused by non-standard font and poor performance of classifier [1], which shows the important role of segmentation in recognition. In recent years, to improve the recognition rate of handwritten text, people have turned their focus of study to character segmentation technology.

The first algorithm that was widely used to segment Chinese characters is projection histogram [2]. This kind of algorithm is used to segment printed characters, but it also can segment handwritten characters effectively which written neatly and have wider space. Connected components analysis [3] can segment overlapping characters and tilted characters well, but it may cause connected components smashed excessively, making badly broken text image hard to re-merge. And real touching characters cannot be segmented by connected

ISSN: 1975-0080 IJMUE

components, they need to be re-segmented by adding touching character template in subsequent recognition module or by other means. Literature [4] presents a kind of dynamic programming algorithm to search optimal path. It is based on recognized character segmentation algorithm and directs the segmentation by recognizing module. The result of recognition plays a crucial role in segmentation, and segmentation is the by-product of recognition. Segmentation result of this kind of algorithm depends on performance of classifier, and execution efficiency is low as characters need to be recognized repeatedly. It is the development direction of segmentation algorithm, because it is most closely to human brain's reaction to characters. All above-mentioned segmentation proposals have their advantages, but their accuracy rates are still not high because of the arbitrariness of handwritten text. The recognition rate of handwritten text is also in a low level, so there is still a long way to go. This paper studies non-defining Chinese handwritten text to mainly discuss relevant technology and algorithms in its segmentation.

2. Image Pre-processing

Images studied in this paper are handwritten text images and obtained by scanning equipment. They are 8-bit grey level images, and their format is uncompressed BMP. Interferences may come from micro contamination in text background, ink dot, broken points and different stroke kerning or performance of equipment and other reasons. This paper employs binarization processing and smooth denoising to eliminate micro contamination in original image and noise generated in process of scanning.

2.1. Smooth Denoising

Smooth filter is to eliminate isolated noise spots and to fill blank spots of target area so that burrs and nicks on character edge line can be reduced for the benefit of subsequent algorithm. Smoothing in this paper is realized with mean filter. The size of mask is determined by the size of object which will be integrated into the background. Elimination of noise spots is realized by smooth algorithm of mask. As shown in Figure 1.

(a) is the original image, there are some micro noises which are smaller than stroke. (b) is the image after filtering, obviously, the noises have integrated into the background.

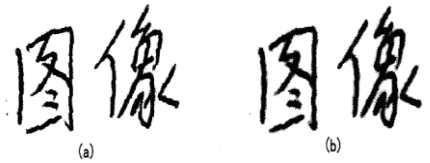

Figure 1. Comparison on Images Before and After Filtering

2.2. Binaryzation of Images

As shown in Figure 2, this paper adopts standard Ostu algorithm binaryzation text images. As Figure 2 shown, binaryzation image without smoothing has noise in the background, and characters' stroke also has burrs.

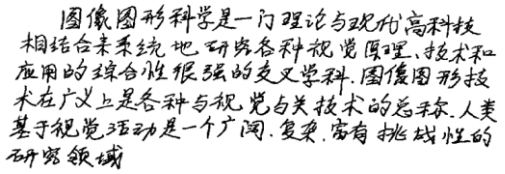

Figure 2. Text Image After Binaryzation

2.3. Estimation of Stroke Width

Calculate black pixel's run-length of each row or each column by horizontal scanning and vertical scanning the image and draw run-length histogram as shown in Figure 3. The value calculated in Figure 3 is about 4.5. After getting the estimate value of ASW, according to measurement of experiment, when the size of mask is no more than the minimum odd number of stroke width, under the condition of keeping character's information, filter of noise can get the best result. This paper uses mean value of two peak values as estimated value of ASW. That is:

$$ASW = \frac{sw1 + sw2}{2} \tag{1}$$

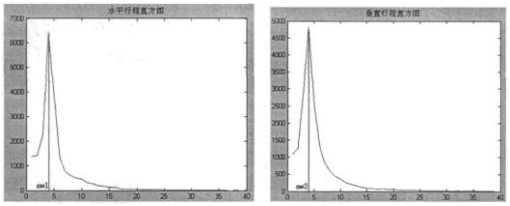

Figure 3. Run-length Histogram of Binaryzation Text Image

3. Extract of Character Row in Text

The process of segmenting multi-row handwritten text into single characters usually is divided into two steps. The first step is segmenting the text into rows to extract character row. The second step is segmenting character rows into single characters. In order to extract character row from text accurately, domestic researchers have put forward many relevant algorithms. All these algorithms can achieve certain effect on extract, but they all have some disadvantages. In this paper presents a kind of multi-step extract algorithm for searching nonlinear row. Details are as follows:

Step1: To find the entrance between character rows. We take left part of the image to perform horizontal projection. As shown in Figure 4, at the beginning of each row, the image is not overlapping, and rows can be segmented easily by simple horizontal projection.

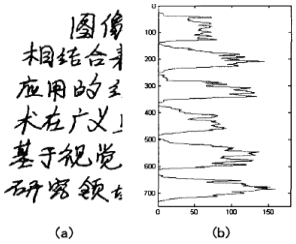

Figure 4. Text Image and Its Horizontal Projection of Left Part

Step2: To search row segmentation line. To search segmentation line is to search each segmentation point on each segmentation line. The first segmentation point of each row segmentation line is located in the first column. The last segmentation point of each segmentation line is located in the last column of image. The searching process will be as follows:

Put point p1 as the first segmentation point to search in three directions. The search continues when meets white pixel points and pauses when meets black pixel points or up and down boundary points of image, and record the termination points of g1, g2 and g3. If there is no black pixel point during searching until the last column, then the last column is recorded as segmentation point directly, and the search of this row is fulfilled.

Calculate the step length l1, l2 and l3 of the first step search. The definition of step length is column numbers advanced from starting point p1 to termination points of g1, g2 or g3.

Compare the size of step length. When comparing, as we hope segmentation line is as horizontal as possible, so l1 is weighted, that is, we will compare al1, l2 and l3. " a" will take the number more than 1. Determine the right direction of the search according to the size of step length, and the right direction is one whose step length is the longest.

Put the central point of line segment which from p1to g1 as the second segmentation point, and record its coordinate p2 (x2,y2).

Search next segmentation point p3 with p2 as the starting point to last column, and record the last segmentation point.

Repeat above steps, search other segmentation points of each row.

Step3: To extract row. The first row is located between up boundary line of the text and the first segmentation line, the last row is located between down boundary line of the text and the last segmentation line, and other rows are located between each two segmentation lines. All character rows are extracted by this way. Character rows extracted by this method preserve the original information of character furthest. Even characters are a little slant, they will be extracted correctly.

4. Segmentation of Characters

4.1. Segmentation of Non-touching Characters

In order to improve accuracy rate of algorithm, this paper adopts algorithm presented in literature [5] for direct modeling. Specific approach is as follows: assume that character row image to be segmented is Im×n, so the whole text can be indicated by $I = \{v_{i,j} | 1 \le i \le m, \ 1 \le j \le n\}$. Text image at this moment can be seen as a top-down directed graph with m layers, each pixel is corresponding to a nodal point in the graph. Each node-set constitutes a layer, between layers is full connection. Set parameters of model as follows:

Observation probability bk(j) of nodal point Vkj:

$$b_k(j) = Pixel_{k,j}, \quad 1 \le k \le m \quad 1 \le j \le n \tag{2}$$

4.1.1. Probabilities of Initial State and State of Termination: As starting point and terminating point of segmentation path can be in first layer or the last layer, probabilities of initial state and state of termination set here both are 1.

4.1.2. Probability of State Transition aij: As in practice, it is unnecessary to make a full transition. In order to segment characters which have more serious overlapping, this paper takes five transition forms. In multi-layer directed graph, we need to find an optimal one. Use to record the highest probability which obtained when along a certain path reaching k grid

node of j layer; use to record the nodal point which makes grid nodes of j layer in j-1 layer achieve the highest probability. The details are as follows:

Calculate the nodal point probability of the first layer.

$$\delta_1(k) = \pi_k b_1(k), \quad 1 \le k \le n, \quad \varphi_1(k) = 0 \tag{3}$$

Calculate the highest nodal point probability of the second layer to m-1 layer, and optimal nodal point of previous layer.

$$\delta_{j+1}(k) = \max_{k-2 \le i \le k+2} [\delta_j(i) a_{i,k} b_{j+1}(k)], \quad 1 \le j \le m-2, \quad 1 \le k \le n \tag{4}$$

Calculate the highest nodal point probability of m layer, and optimal nodal point of previous layer.

$$\delta_m(k) = \max_{k-2 \le i \le k+2} [\delta_{m-1}(i) a_{i,k} b_m(k) \gamma_k], \quad 1 \le k \le n \tag{5}$$

According to Viterbi algorithm, paths obtained by backtracking from m layer to the first layer constitute all segmentation paths of original character string.

4.2. Segmentation of Touching Chinese Characters

After Viterbi algorithm, Chinese characters have many segmentation paths. These paths segment rows into single small character blocks, and we call these small blocks components. Given that touching characters are usually wider than non-touching characters, we use width discrimination method to decide touching characters.

Extract all components and calculate height hi and width wi of each component.

Estimate average width WE of Chinese characters: Sort component's height hi, statistically, value of Hm in sequence can indicate the average height of Chinese character, after that, the average height can be measured in experiment according to features of Chinese character.

For each candidate component, if $w_i > c \cdot w_E$, the candidate character is regarded as touching character which needs segmenting touch area.

The key for segmenting touch area is to find the location where touching occurs. The form of touching usually is interconnection or intersection. This paper adopts stroke analysis presented by literature [6] and Viterbi algorithm to segment touching Chinese character.

Step1: detection of feature points and extract of stroke. Chinese characters are thinned first, after that, they receive feature points detection. Feature points include end points, fork points and corner points.

Step2: detection of touching points. Touching often occurs on long stroke in central part of image because of the writing habit of Chinese character, so search is just limited in central part of image, and search long stroke first. The feature point found is touching point.

Step3: segmentation. After confirm the touching point, segment the touching character from touching point. Segmentation cannot be performed in vertical path, because top and down area of touching area may have overlapping, therefore, segmentation must be performed in a proper path. In order to keep the consistency of segmentation path in touching characters and non-touching characters, this paper still use Viterbi algorithm to perform nonlinear segmentation after separation. The process of segmenting touching characters is shown in Figure 5.

Figure 5. Process of Segmenting Touching Characters

After segmentation of touching characters, put generated segmentation paths into candidate paths, constituting set of all candidate paths shown in Figure 6. All candidate paths segment character rows into thinner components which can form into characters through merging. The process of merging components is seen as optimization to all candidate paths in this paper, that is to say, preserve correct paths and delete wrong paths. The rest paths will segment characters from rows.

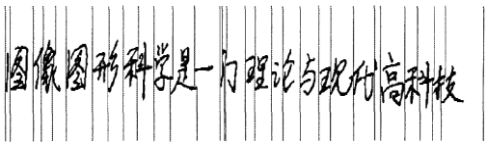

Figure 6. Set of All Candidate Paths

5. Optimization of Segmentation Paths Based on Genetic Algorithm

After segmentation, most of overlapping and touching Chinese characters can be segmented correctly, but wrong segmentation still occurs, the reason for this is mainly due to excessive segmentation. Excessive segmentation often occurs on left-right structure Chinese characters. Because some Chinese character components themselves can be characters, therefore, previous candidate paths must be optimized, searching optimal paths and merging some components to achieve correct segmentation result. This paper will employ genetic algorithm to optimize segmentation path. The structure of genetic algorithm is as follows:

5.1. Encoding

Logical form in this paper is indicated by binary 1, 0. Each segmentation path is identified as 1 or 0, and "1" means preserving, "0" means deleting. The first situation shall be considered is no punctuation in character row: there are 31 candidate segmentation paths in Figure 7, and they can be indicated by 31-bit gene. As the first path and the last path are preserved definitely, so the corresponding gene bit must be 1, and it does not involve into computing. Therefore, actual code number of gene is 29 bits [11111111111111111111111111111] which less 2 bits than original bits.

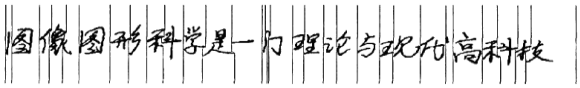

Figure 7. All Candidate Segmentation Paths

Figure 8 describes the corresponding path that selected by a random gene [10111000010101111100100001000]. For the sake of convenience, paths that should be deleted are indicated by dotted line.

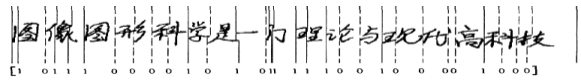

Figure 8. Corresponding Path That Selected by a Random Gene

If character row contains punctuation, besides preserving the first path and the last path, we also need preserve the left and right paths of the punctuation as shown in Figure 9, and there are 31 candidate segmentation paths altogether.

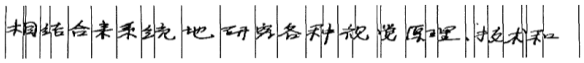

Figure 9. Character Row Containing Punctuation and Segmentation Paths

The first and the last paths should be preserved, and the 24th, 25th paths also should be preserved, because punctuation is contained between them. Therefore, all segmentation paths can be indicated only by 27 bits code numbers. The corresponding path that selected by a random gene [011010110111011101101111011] is as shown in Figure 10.

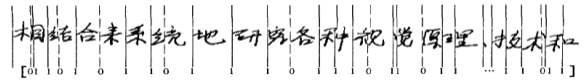

Figure 10. Corresponding Path That Selected by a Random Gene

5.2. Parameter Setting

Set of population size: Given that the request of computing speed and accuracy rate, the size of population in this paper is set to 40.

Evolution algebra: We usually not only request the convergence of the system, but also request a higher quality of solution. As fitness in this system is obtained indirectly, the calculation is complicated relatively. Therefore, the evolution algebra is not suitable too big. The biggest evolution algebra set in this paper is 40.

Selection operator selection: This paper adopts rank-based fitness assignment to select, and pressure is 2, generation gap is 0.9.

Crossover operator selection: According to features of code numbers in this paper, we select single-point crossover of genetic algorithm to reduce destructiveness to individuals. In order to improve rate of convergence, the paper employs a simple strategy of storing optimal individual, that is to say, using binary crossover rate. Use low crossover probability for individuals with larger convergence, on the contrary, use high crossover probability for individuals with smaller convergence. Calculate crossover probability according to the following formula (8). “Pc1, Pc2” of this paper are set to 0.4 and 0.7 respectively.

$$P_c = \begin{cases} P_{c1} & f \geq f_{avg} \\ P_{c2} & f < f_{avg} \end{cases} \quad (8)$$

5.3. Experiment

Through above-mentioned processing, all candidate paths in Figure 5.10 are optimized into ones as shown in Figure 5.15(a). Extract components generated in segmenting paths by means of bounding rectangle, and single characters obtained are shown as Figure 5.15(b). All work to segment characters of recognition system is finished.

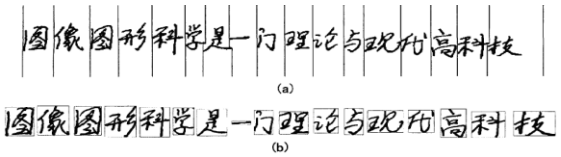

Figure 11. Results of Segmentation Paths

6. Conclusion

In the decades of development of Chinese character recognition technology Chinese character recognition experienced from the letters to the word, from word to text, from printed to handwritten. This paper studies the Chinese characters handwritten character segmentation. After image preprocessing, compared the characteristics and limitations of the character line extraction algorithm, proposed a multi-step search for the non-linear lines of characters extraction algorithm and non-adhesions and adhesions character segmentation method. Through optimization of the Viterbi algorithm, so that the segmentation of the non-adhesion character is more effective, and it can be divided to overlap more serious characters. For adhesion character segmentation, first stroke analysis method to identify feature points, and then find the sticking point with the Viterbi algorithm to generate a split path. Experiments show that this algorithm is accurate and effective, segmentation for touching character and adhesion character achieved good results. Design genetic algorithm which optimizes the candidate path, using the average character forming probability as the fitness function. Design the preservation and deletion of the marking path of the logic code gene. Select the right parameter to void poor convergence or local maximum. The test showed that the genetic algorithm of this paper performances efficiently and searches rightly.

Acknowledgments

This work supported by the Key Program of National Natural Science Foundation of Heilongjiang No.ZD201309, and Project supported by the Maor International Joint Research Program of China (Grant No. 2013DFA71120).

References

[1] L. Qingzhong, "The Research of Character Segmentation in OCR and Text Extract", Tingjing: Nankai University, **(2001)**.
[2] L. Y, "Machine Printed Character Segmentation — An Overview", Pattern Recognition, vol. 28, no. l, **(1995)**, pp. 67-80.
[3] S. Jie and C. Yu, "A Survey of Methods in Handwritten Chinese Character Segmentation", Computer Technology and Development, vol. 16, no. 6, **(2006)**, pp. 184-190.
[4] M. Rui, "Research on Segmentation of Unconstrained Handwritten Characters", Nanjing University of Science and Technology, **(2007)**.
[5] Z. Lian and P. Shi, "A meta-synthetic approach for segmenting handwritten Chinese character strings", Pattern Recognition Letters, vol. 26, **(2007)**, pp. 1498-1511.
[6] Z. Shuyan, G. Jie and S. Pengfei, "Segmentation of Connected Handwritten Chinese Characters Based on Stroke Analysis and Background Thinning", Journal of Shanghai Jiaotong University, vol. 37, no. 9, **(2003)**, pp. 1434-1437.

