A Model of Stroke Extraction from Chinese Character Images

Ruini Cao, Chew Lim Tan
School of Computing, National University of Singapore
Email: caoruini, tancl@comp.nus.edu.sg

Abstract

Given the large number and complexity of Chinese characters, pattern matching based on structural decomposition and analysis is believed to be necessary and essential to off-line character recognition. This paper proposes a new model of stroke extraction for Chinese characters. One problem for stroke extraction is how to extract primary strokes. Another major problem is to solve the segmentation ambiguities at intersection points. We use the degree information and the stroke continuation property to tackle these two problems. The proposed model can be used to extract strokes from both printed and handwritten character images.

1. Introduction

Extracting the stroke information is important in off-line character recognition. Given the large number and complexity of Chinese characters, pattern matching based on structural decomposition and analysis is believed to be necessary and essential. [1-4] Recently, there is a growing interest in obtaining temporal information from static line images to improve the overall performance of the recognition system. [5-6] This approach is considered as a bridge from the off-line handwriting character recognition problem to the on-line one that is generally agreed to have better performance. Stroke segmentation is however one of the prerequisites of extracting dynamic writing information.

There are two major problems for stroke extraction. One is how to extract primary strokes. Another problem is to solve the segmentation ambiguities at intersection points. Most existing methods for stroke extraction use thinning process. [3-7] These approaches have an intrinsic problem of spurious branches or pattern distortion, which may lead to unreliable extraction of strokes. Algorithms without thinning process exploit other kinds of stroke information, such as stroke width variations, curvature changes or stroke continuation property. [1,8-10]

In this paper, we propose a new model that combines the stroke continuation property with the component connectivity information. The proposed model is entirely based on thick-line images. Compared with the model in [10], the proposed model exhibits better performance in solving the ambiguities and grouping broken strokes. We propose a different way to estimate the line tangent orientation. This way proves to be more discriminating. We also come up with a general way to group non-smooth zigzag strokes.

The remainder of the paper is organized as follows: In section 2 to section 4, we explain how to extract primary strokes, how to separate overlapped strokes, and how to extract complete strokes. In section 5, we present some experiment results and discussions. Section 6 gives the conclusion.

2. Extraction of primary strokes

We define the primary stroke in terms of the degree. The degree of a pixel is the number of the branches incident on it. According to the degree information, a line image can be divided into three regions: end point regions, regular regions and singular regions. A pixel belongs to an end point region if its degree is equal to 1. If the degree of a pixel is equal to 2, it belongs to a regular region. Singular regions are made up of pixels whose degree is 3 or more. (See Figure 1) The primary stroke is the connected end point regions and regular regions.

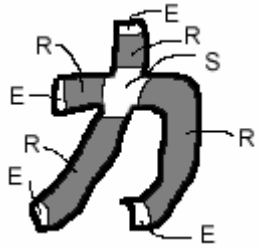


Figure 1. Decomposition of a line picture into regular regions (R), end point regions (E), and a singular region (S)

We use the direction contribution of the segments to estimate the degree for each pixel on the thick-line image. Given a binary image, for each black pixel $p(r,c)$, let $D_k(r,c)$ denote the orientation distance between the pixel and the boundary point along the kth quantized orientation, where $k = 1, 2, \ldots, M$, and M is an integer that

denotes the quantization number from 0^o to 360^o. The orientation distance is called point-to-boundary orientation distance (PBOD).

The distribution of the PBODs contains information about the degree of each pixel. We can easily tell that the PBODs along the quantized orientations of the branches are much larger than that along other quantized orientation. So, to estimate the degree of each pixel, we need only calculate the number of the crests of the distribution of all the PBODs of the pixel. Figure 2 shows some cases of the distribution. The resolution of quantization is 3^o.

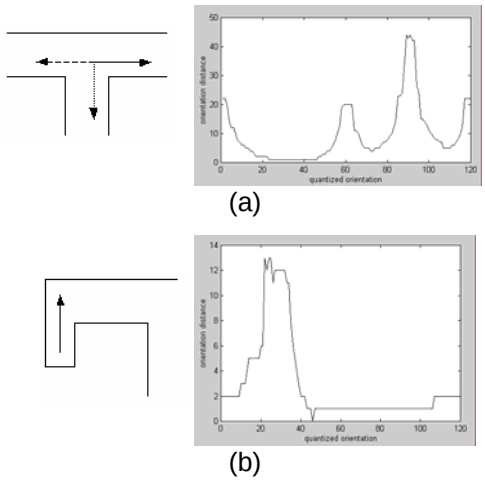


Figure 2. Distribution of PBODs: (a) Pixel at the singular region; (b) Pixel at the end point region

3. Separation of overlapped strokes

To separate overlapped strokes, we need to interpret the singular region in a correct way. We construct a 3-D ρ-space to tackle this problem.

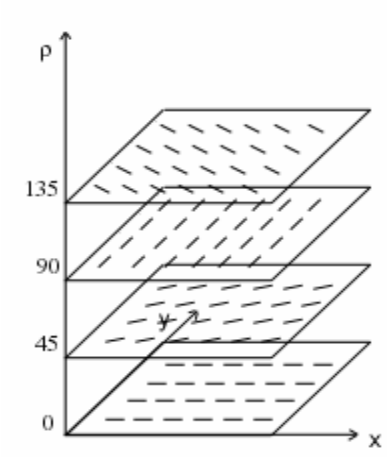


Figure 3. Illustration of the ρ-space

The ρ-space is a 3-D space where the first 2 dimensions are the spatial image dimensions, and the third dimension is the orientation dimension (See Figure 3). Each element *(i, j, m)* in the ρ-space is set to be 1 if and only if the tangent orientation of the line passing the pixel *(i, j)* is *m*; otherwise, it is set to be 0. We can imagine that there are *N* orientation planes in the ρ-space and the 2-D image is mapped to the ρ-space by putting each pixel on some of the planes depending on the tangent orientations of the lines passing the pixel.

The ρ-space helps extract smooth stroke segments. Consider a pixel *p(i, j, k)*. It has 24 neighbors, counting the points on the plane *(k-1)* and the plane *(k+1)*. (Note that the first plane and the last plane are contiguous.) A transition between adjacent planes does not violate the smoothness of a curve since the tangent orientation changes only a little. Hence, finding the connected components in the ρ-space extracts the smooth strokes.

The ρ-space has the advantage of being able to represent more than a single orientation at a single point. For a pixel in the end point region and the regular region, there is only one line passing it, so the pixel is mapped only onto one plane in the ρ-space. Whereas, for a pixel located at the singular region, there are at least two lines passing it. Obviously, the tangent orientations of the two lines are different. So, the pixel will be mapped onto two different planes in the ρ-space. The two overlapped lines passing the pixel will then be separate from each other. (See Figure 4)

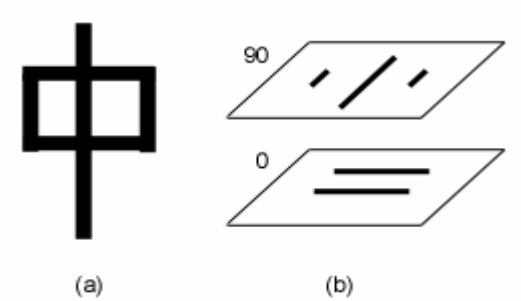


Figure 4. Segmentation in the ρ-space: (a) A character image; (b) Overlapped strokes are mapped into different planes and are separated

Now, the key problem is how to find the tangent orientations of the lines passing each pixel. We also exploit the direction contribution of the segments.

Given a binary image, for each black pixel *p(r,c)*, let $D_k'(r,c)$ denote the orientation distance between two boundary points along the *k*th quantized orientation, where $k = 1, 2, \ldots, N$, and *N* is an integer that denotes the quantization number from 0^o to 180^o. The orientation distance is called boundary-to-boundary orientation distance (BBOD).

We find all the crests of the distribution of the BBODs of one pixel and take the quantized orientations corresponding to the crests as the tangent orientations of the lines passing this pixel. This is based on the observation that the BBOD along the line tangent orientation is often much larger than the BBODs along other orientations.

The tangent orientation obtained in this way is more discriminating than that obtained in [10] where the tangent orientation (called OLLS there) is obtained by binarizing the BBODs using a threshold defined as the mean of all the BBODs of the pixel. First, the latter way may miss some short strokes joining with long strokes since the BBOD along the orientation of the short stroke may be less than the mean value. Secondly, binarization of BBOD may cause confusion at two branches when the trough between them is larger than the mean value that usually occurs near junctions.

4. Extraction of complete strokes

We summarize the proposed model in the following.

Stage 1. According to the line tangent orientation, we first map the 2-D line image into the 3-D ρ-space. Then we perform connected components labeling in the ρ-space. After projecting each 3-D connected component onto a 2-D plane, we get the smooth stroke segments. One assumption here is that overlapped strokes are smooth at the intersections.

Stage 2. The strokes obtained so far are usually over-segmented due to L-type connection or bad line smoothness. We further group the segments by linking them at the pixels where primary strokes are broken.

In this paper, we only address the stroke segmentation problem. As for obtaining dynamic writing information, after we get segmented strokes, we can utilize some popular human writing rules to determine the most possible writing direction of strokes. [6] For example, Chinese characters usually follow the top-to-bottom and left-to-right rules.

5. Experiments and discussions

The proposed model was tested using a set of printed Chinese characters that are selected randomly. The number of the characters in the testing set is 111. There are about 849 strokes in total. The model segmented 806 strokes correctly. The accuracy is 93%. There are about 98 characters whose strokes are all segmented correctly. The correct rate is 88%. Among the correctly segmented characters, the complexity is about 7 strokes per character on the average.

Some of the typical cases are shown in Figure 5 to Figure 7. The results by the model in [10] are also shown to give some comparison. We can easily see that the proposed model performs better in separating overlapped strokes (Figure 6) and preserving the connectivity of primary strokes (Figure 6 and 7). The strokes extracted by the proposed model are closer to the correct way of segmenting Chinese character strokes. Figure 8 shows an example of extracting strokes from handwritten character images. We can see that so long as the overlapped strokes are smooth at intersections, the proposed model can separate them successfully.

(a) The character image

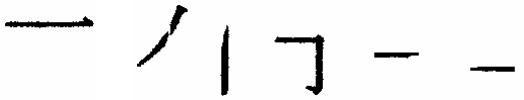
(b) Extracted lines using the model in [10]

(c) Extracted strokes by the proposed model

Figure 5. Segmentation example: the two models produce the same results

(a) The character image

(b) Extracted lines using the model in [10]



(c) Extracted strokes by the proposed model

Figure 6. Segmentation example: the proposed model is more discriminating

During the test, one important parameter is the resolution of quantization. If the resolution is high, the orientation selectivity is high. That means it can separate overlapped strokes intersecting at smaller angles. However, if the resolution is too high, in the case that one of the overlapped strokes has a high degree of curvature at the intersection, the estimated orientations will be not continuous. That means the smoothness information is lost. Therefore, there is a tradeoff in selecting the resolution of quantization. When we did the experiments, the resolution of quantization was selected at 3^{o}.

It is necessary to mention here that the strokes are not completely equivalent to the strokes defined as the path between pen-up and pen-down in the on-line character recognition system. This is one intrinsic difference between off-line and on-line system. Namely, on-line system traces pen-up and pen-down, whereas off-line system finds two ends of one line. When two strokes are linked end to end, they cannot be segmented from each other in the off-line case (See the first connected component in Figure 7c).

6. Conclusion

A new model of stroke extraction is proposed in this paper. First we extract the primary strokes according to the degree information of each pixel on the thick-line image. Then we construct a 3-D ρ-space to solve the segmentation ambiguities at intersection points.

The proposed model is performed entirely on thick-line character images, so no distortions are introduced. It also exhibits better performance in separating overlapped strokes and preserving connectivity of primary strokes. This model can be used to extract strokes from both printed and handwritten character images.

7. References

[1] F. Chang, Y. Chen, H. Don, W. Hsu and C. Kao, Stroke segmentation as a basis for structural matching of Chinese characters, *Proceedings of 2nd International Conference on Document Analysis and Recognition*, Japan, 1993

[2] W. Ip, K. Chung and D. Yeung, Offline handwritten Chinese character recognition via radical extraction and recognition, *Proceedings of 4th International Conference on Document Analysis and Recognition*, Germany, 1997

[3] K. Liu, Y. S. Huang and C. Y. Suen, Robust stroke segmentation method for handwritten Chinese character recognition, *Proceedings of 4th International Conference on Document Analysis and Recognition*, Germany, 1997

[4] H. Chiu and D. Tseng, A novel stroke-based feature extraction for handwritten Chinese character recognition, *Pattern Recognition*, **32**, 1999, pp. 1947-1959

[5] Y. Kato and M. Yasuhara, Recovery of drawing order from scanned images of multi-stroke handwriting, *Proceedings of 5th International Conference on Document Analysis and Recognition*, India, 1999

[6] J. Zou and H. Yan, Extracting strokes from static line images based on selective searching, *Pattern Recognition*, **32**, 1999, pp. 935-946

[7] Y. Nakajima, S. Mori, S. Takegami and S. Sato, Global methods for stroke segmentation, *International Journal on Document Analysis and Recognition* **2**, 1999, pp. 19-23

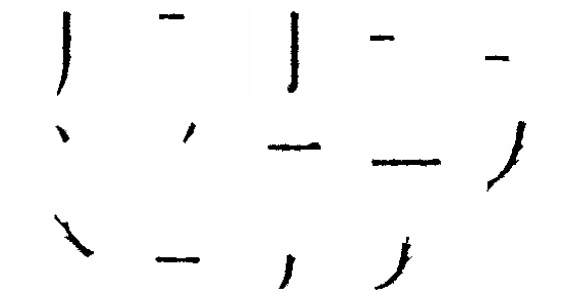
[8] L. Y. Tseng, C. T. Chuang, An efficient knowledge-based stroke extraction method for multi-font Chinese charaters, *Pattern Recognition* **25**, 1992, pp. 1445-1458

[9] C. M. Privitera, R. Plamondon, A system for scanning and segmenting cursive handwritten works into basic strokes, *Proceeding of 3rd International Conference on Document Analysis and Recognition*, Canada, 1995

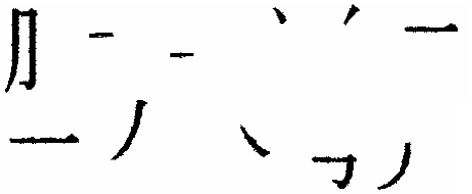
[10] Y. S. Chen and W. H. Hsu, An interpretive model of line continuation in human visual perception, *Pattern Recognition* **22**, 1989, pp. 619-639

[11] S. Liang, M. Ahmadi and M. Shridhar, Segmentation of handwritten interference marks using multiple directional stroke planes and reformalized morphological approach, *Transactions on Image Processing* **6**, No. 8, 1997, pp.1195-1202

(a) The character image



(b) Extracted lines using the model in [10]



(c) Extracted strokes by the proposed model

Figure 7. Segmentation example: the proposed model groups zigzag strokes

(a) The character image

(b) Extracted strokes

Figure 8. Segmentation results for handwritten characters