行政院國家科學委員會專題研究計畫成果報告

數位圖書館及博物館之自動化資訊處理(3/3)—子計畫二：

中國碑帖文字之擷取、辨識與儲存

Segmentation and Recognition of Chinese Characters in Cursive Script in Calligraphy Documents

計畫編號：NSC 90-2213-E-009-048
執行期限：90 年 8 月 1 日至 91 年 7 月 31 日
主持人：李錫堅　　國立交通大學資訊工程系

一、中文摘要

書法是我們中國國粹之一，而草書是變化極為複雜的一種字型。在這計畫中，我們設計了一個草書書法字帖的文字自動切割與辨識系統。如此一來，便能夠將草書書法自完整保存下來。系統的輸入影像為二值化後的書法字帖點陣圖像。文字切割和文字是別為此系統中的兩個主要的模組。在文字切割模組方面，我們對輸入影像建構了一個最短距離的對應圖，這個對應圖紀錄了影像中每一個點的最短路徑。接下來利用這個對應圖並配合垂直投影將可以找到垂直行文字的切割路徑。再同樣的對每一個垂直行文字建構最短距離的對應圖找到初始的水平文字切割路徑，最後利用限制切割路徑和中文草書書法自的特性來去除掉多餘的水平文字切割路徑。在文字辨識模組方面，我們希望找到合適草書書法自使用的文字辨識核心。考慮四種不同的統計式文字特徵抽取方法：邊緣方向數、通過筆劃數、 Oka’s cellular 特徵和 peripheral background area 特徵。利用這四種文字特徵抽取方式的各種不同權重的組合和五種計算特徵距離的方法，找出對於書法影像有最高辨識率的組合。實驗影像是由五位中國古代書法家的草書作品中選出的五十五張字帖影像。在垂直行文字的切割成功率有 98.23%，而在水平文字切割的正確率為 84.06%。

Abstract

The calligraphy is one of the quintessence of Chinese culture. The Chinese cursive script is a quite complicated style in calligraphy script styles. In this project, we design an automatic segmentation and recognition system for Chinese characters in cursive script. Thus, we can preserve the Chinese characters in cursive script in a database. The input of our system is binary Chinese cursive script calligraphy image without noises. Our system contains two major modules: characters segmentation and characters recognition. In the characters segmentation module, we first construct a shortest distance map that contains each shortest path for each point of the input image. Then combine the shortest distance map with the vertical projection to find the vertical text line segmentation paths. Next, we apply the shortest distance map in each text line to obtain initial horizontal character segmentation paths. Finally, reduce the horizontal character segmentation paths by using the path constraints and cursive script features. In the characters recognition module, in order to find a good OCR engine that has a high recognition rate for Chinese characters in cursive script, we consider four statistical feature extraction methods: contour directional features, crossing count features, Oka's cellular features and Peripheral background area features. We combine these four feature extraction methods with five

feature distance measurements methods to select our OCR kernel with a highest recognition rate of our testing characters in cursive script.

Keywords: cursive script , calligraphy, segmentation, recognition

1. Introduction

Cursive script originated in the second century B.C. as an abbreviated form of clerical script. It is matured in the third and fourth centuries A.D., culminating in the calligraphy of Wang Hsi-chih(303-361). Cursive script is a shorthand form used for personal notes and letters. Therefore, touched characters appear frequently in cursive script. The qualities of cursive script vary with the freedom of brush and ink and large variations in handwriting of different authors. In order to preserve the Chinese characters in cursive script, three problems will be concerned. They are segmentation of an irregular calligraphy image and connected Chinese characters in cursive script and recognition of Chinese characters in cursive script. The system proposed here contains two major tasks: segmentation and recognition.

For image segmentation, several methods have been proposed in the literature. Lu[1] presented and overview of machine printed character segmentation. Lu and Shridhar[2] also reviewed handprinted word, handwritten numeral and cursive word segmentations. Casey and Lecolinet[3] classified character segmentation methods into three strategies: (1)dissection methods, (2)holistic methods, and (3)recognition-based methods.

Two major approaches locating possible segmentation positions are projection profile analysis and connected-component extraction. The former approach projects all black pixels in the X-axis(or Y-axis) and selects positions whose numbers of accumulative pixels are smaller than a threshold as possible segmentation positions. The latter one merges neighboring connected runs that consist of continuous black pixels to be connected components. The boundaries of these components are considered as possible segmentation positions. However, these methods are limited in that they can only detect “linear” segmentation paths. Wang and Jean[5] derived a non-linear cutting path from touching characters by identify the shortest path. Tseng and Lee[4] applied a probabilistic Viterbi algorithm to obtain non-linear possible segmentation paths from an entire text line. To reduce the number of possible segmentation paths, heuristic checking rules are employed to remove redundant segmentation paths. In the system, we first segment the input calligraphy image into vertical text line images. Then, we segment horizontal character segmentation in each vertical text line into Chinese characters in cursive script images. Finally we recognize the character images with our cursive script OCR kernel.

2. Vertical Text Line Segmentation

2.1 Segmentation Path and Shortest Distance Map

In vertical segmentation, a segmentation path is a set of consecutive points and connects two points in both top and bottom edges of the image. A text line is the area between two segmentation paths. Figure 2.1.1 shows the imaginative vertical text line segmentation result of a Chinese cursive script image.

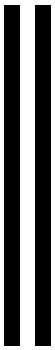

Figure 2.1.1 Imaginative segmentation path and text line

2.1.1 Vertical segmentation paths

Because of the large variations in Chinese cursive script handwritings, the vertical segmentation paths (Figure 2.1.2) must conform to the following three demands: (1) the direction of a segmentation path is 4-way; (2) the cost defined on a segmentation path is less than all paths which have the same start point and end point (Shortest path) ; (3) Down-ward first.

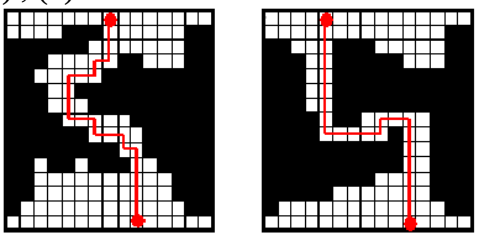

Figure 2.1.2 Vertical segmentation path

2.1.2 Shortest distance map

Here, we design a shortest distance map to obtain the shortest path of all points in the image to some given points.

Def. 1: A starting line is and edge of the given image for the beginning of the segmentation paths.

Def. 2: An ending line is the opposite edge of the starting line in the image.

Def. 3: Starting points of the segmentation paths are all points on the starting line.

Def. 4: A city-block distance is the coordinate difference of a path movement from a point *(x, y)* to its neighboring point, {*(x+1, y), (x-1, y), (x, y+1), (x, y-1)*}.

Def. 5: A shortest path to a given point *(x, y)* is a path from a starting point to the point to the point that has the minimum sum of city-block distances. Let $D_{x,y}$ denote the minimum sum of city-block distances of the point *(x, y)*.

Figure 2.1.3 An example of shortest distance map

2.2 Variants of the Shortest Distance Map

In figure 2.2.1, Paths A, B, C and D are possible vertical segmentation paths. But in observation, Paths A and C are imperfect vertical segmentation paths because they pass through characters. Paths B and D are the wanted vertical segmentation paths.

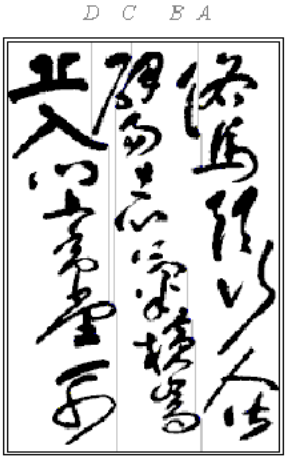

Figure 2.2.1 Possible vertical segmentation path

2.2.1 A property of Chinese characters

If a segmentation path passes by the right side of a small radical of a character, i.e., passes through the character, it may turn left in its following moves. We design a path finding algorithm that can avoid this situation. Therefore, we define weights on city-block distances to make vertical segmentation paths right-turn more easily than left-turn.

Def. 7: The weighted distance between a point and one of its 4-neighbor is determined by the color, black or white, of the point and the location of the neighbor, top, bottom, left or right, of the neighbor.

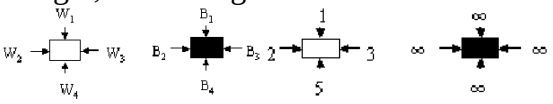

Figure 2.2.2 weighted distance and the experimental result

2.2.2 Shortest path back tracing

In figure 2.2.3, PathA and PathB are ina tie condition in a part of shortest distance map, but only one path is required for the segmentation path. The breadth-first search (BFS) is adopted to construct the shortest distance map.

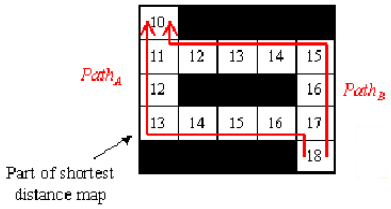

Figure 2.2.3 The tie condition of two path

2.3 Vertical Text Line Segmentation Path

2.3.1 Initial segmentation paths

The initial segmentation paths are selected with a threshold, V_{thr}:

$V_{thr} = \sum_{x=0}^{w} \frac{V(x)}{2w}$, where w is the image width. If two initial segmentation paths are very close, we choose the one that has the smaller weighted city-block distance $D_{i,h}$ in the distance map, where h is the height of the imag.

2.3.2 Segmentation path refinement

The following condition is used. (1) The starting line is the top edge of the input image. (2) Set the weighted distance *B1~B4*, *W1~W4* according to our experimental result.

2.3.3 Error correction

If a component c is small and very close to a segmentation path, we reassign this component c to the text line that has a nearest connected component with the component c.

3. Horizontal Character Segmentation

The horizontal segmentation paths must satisfy to the following four demands: (1) the direction of a segmentation path is 4-way, (2) the shortest paths from points in the left edge of the image, (3) right-ward first, (4) black points crossed.

3.1.1 Initial Horizontal Character Segmentation Path

We train the weighted distance to make the segmentation path cut the stroke which connects two characters from left-up to right-down.

3.1.2 The weighted shortest distance map of all points in a text line

The following conditions are introduced: (1) the starting line is the left edge of the text line, (2) set the weighted distances *B1~B4*, *W1~W3* according to our experimental result. Then, for each point *(w, y)* at the rightmost column of the weighted shortest distance map, we can find a shortest path by tracing the backward edges from the point.

3.2 Candidates Selection and Reduction

We use two strategies to determine the horizontal character segmentation path.

1. Select candidates according to path constraints.
2. Reduce candidates by rules about the features of the Chinese cursive script.

3.2.1 Paths selection using the path constraints

Three constraints is adopted to determine the candidates of horizontal

character segmentation paths.

Constraint 1: Only one path is kept in the gap of two consecutive components.

Constraint 2: Only one path is selected from partially overlapping paths.

Constraint 3: A path cannot cross more than two components.

3.2.2 Paths Reduction

We try to reduce the candidate segmentation paths by using four cursive script features listed below. (1) The last vertical stroke of a character may be very long. (2) Joints between strokes in a character may be very narrow. (3) Small components appear more frequently at the top of a character than at the bottom. (4) Components in the same character cannot separate too far.

Reduction Procedure

1. Assume a candidate path $Path_i$ is selected from the middle of the gap of two consecutive components. If the height of this gap is over a threshold, set $Path_i$ as a confirmative path.
2. For a character section CS_i between candidate paths $Path_i$ and $Path_{i+1}$, if only one of $Path_i$ and $Path_{i+1}$ is a confirmative path, remove the path that is not a confirmative path from the candidate paths.
3. For a character section CS_i between two candidate paths $Path_i$ and $Path_{i+1}$, and none of $Path_i$ ad $Path_{i+1}$ is a confirmative path. Reduce the candidate paths by some conditions.
4. If candidate paths have been reduced in Step 2 or 3, recalculate the information of the character sections and go to Step 2, otherwise stop.

4. Chinese Cursive Script OCR Kernel

4.1.1 Non-uniform segmentation of characters

In order to accommodate handwriting variations, a 2-D character image is first segmented non-uniformly into 8 strips in both the horizontal and vertical directions. These strips have the same numbers of black pixels.

4.1.2 Features

Four features, Contour directional features (CD), Crossing counts features (CC), Oka's cellular features (Cellular) and Peripheral background area (PBA), are adopted in the OCR kernel.

4.2 Measurements of Feature Distance

We evaluate these five measurements of distance between the input character and our training samples of Chinese characters in cursive script. Minimum distance: $d(x, \ll_j) = \sum_{i=1}^{\dim} \left| x_i - \acute{}_{ji} \right|$, where $x=(x_1, x_2,\ldots,x_{\dim})$ be the feature vector of an input pattern and $\ll_j = (\ll_{j1}, \ll_{j2},..., \ll_{j\dim})$ be the mean vector of the reference character j. Other four methods about measure distance are Euclidean distance, Cross correlation distance, Modified Mahalanobis distance and L-Y distance.

4.3 Feature Weight

Feature weights are used to combine feature distance from different feature extraction method.

5. Experimental results

We will evaluate the feasibility of our segmentation and recognition method. The test images contain 55 calligraphy images from fie different ancient Chinese cursive script calligraphers. Table 5.1 shows the segmentation result of each calligrapher's

calligraphy documents. We judge a segmentation failure by grouping one component of a character into another character. The total vertical text line segmentation success rate is 98.23%, and the total horizontal character segmentation success rate is 84.06%.

Table 5.1 Vertical text line segmentation and horizontal character segmentation result

5. References

[1] Y. Lu, • Machine Printed Character Segmentation - an Overview", pattern Recognition, vol. 28, no. 1, pp. 67-80, 1996.

[2] Y. Lu and M. Shridhar, • Character Segmentation in Handwritten Words - an Overview", Pattern Recognition, vol. 29, no. 1, pp. 77-96, 1996.

[3] Richard G. Casey and Eric Lecolinet, • A survey of Methods and Strategies in Character Segmentation", IEEE Trans. Pattern Anal. Mach. Intell., vol.18, no. 7, pp.690-706, 1996.

[4] Y. H. Tseng and H. J. Lee, • Recognition-based handwritten Chinese character segmentation using a probabilistic Viterbi algorithm," Pattern Recognition Letters Volume: 20, Issue: 8, August, 1999, pp. 791-806.

[5] J. Wang and J. Jean, "Segmentation of Merged Characters by Neural Networks and Shortest Path," Pattern Recognition, vol. 27, No. 5, pp. 649-658, 1994.

[6] Xuhong Xiao and Graham Leedham, "Knowledge-based English cursive script segmentation", Pattern Recognition Letters, vol. 21, pp. 945-954, 2000.

[7] L. Y. Tseng and R. C. Chen, "Segmenting handwritten Chinese characters based on heuristic merging of stroke bounding boxes and dynamic programming", Pattern Recognition Letters, Volumn 19, Issue 10, August, 1998, pp. 963-973.

[8] T. F.Li and S. S. Yu, "Handprinted Chinese Character Recognition Using the Probability Distribution Feature," Int. Jour. of pattern Recognition and Artificial Intelligence, vol. 8, no. 5, pp. 1241-1258, 1994.

[9] S. L. Zhao, "Multi-kernel Chinese Character Recognition and A Simplified Language Model Used in General Document Processing Systems," Master thesis, Institute of Computer Science and Information Engineering, National Chiao Tung University, Taiwan, R.O.C., 2000.

[10] F. H. Cheng and W. H. Hsu, "Research on Chinese OCR in Taiwan," International Hournal of Pattern Recognition and Artificial Intelligence, Vol. 5, nos. 1&2, pp. 139-164, June 1991.

[11] Richard G. Casey and Eric Lecolinet, "A Survey of Methods and Strategies in Character Segmentation," IEEE Trans. pattern Anal. Mach. Intell., vol. 18, no. 7, pp. 690-706, 1996.

[12] T. F. Li and S. S. Yu, "Handprinted Chinese character recognition using the probability distribution feature, "Intern. Journal of pattern Recognition and Artificial Intelligence, Vol. 8, no. 5, pp. 1241-1258, 1994.

[13] R. Oka, "Handwritten Chinese-Japanese characters recognition by cellular features," Proc. 6th Intern. Conf. on Pattern Recognition, pp. 783-785, IEEE, October 1991.

[14] C. H. Tung, "A study of handwrittern Chinese text recognition," Ph. D. dissertation, Department of Computer Science and Information Engineering, National Chiao Tung University, Taiwan, R.O.C., 1994.

[15] L. Tu, et al., "Recognition of handprinted Chinese characters by feature matching," Int. Conf. on Computer Processing of Chinese and Oriental Languages, pp. 154-157, 1991.

[16] J. J. Lu, "Shape description of chinese

characters in calligraphy documents," Master thesis, Institute of Computer Science and Information Engineering, National Chiao Tung University, Taiwan, R.O.C., 2000.
[17] C. L. Yeh, "Recognition of English alphabets and numerals in ill-printed name cards," Master thesis, Institute of Computer Science and Information Engineering, National Chiao Tung University, Taiwan, R.O.C., 2000.
[18] R. G. Casey and G. Nagy, "Recursive Segmentation and Classification of Composite Character paterns," Proc. 6th Int. Conf. pattern Recognition (Munich, Germany), 1982, pp. 1023-1026.
[19] H. Fujisawa, Y. nakano, and K. Kurino, "Segmentation Methods for Character Recognition: From Segmentation to Document Structure Analysis," Proc. IEEE, Vol. 80, No. 7, July 1992, pp. 1079-1091.