Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, 19-22 August 2007

RECONSTRUCTING STROKES AND WRITING SEQUENCES FROM CHINESE CHARACTER IMAGES

KAI-TAI TANG, HOWARD LEUNG

Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong
E-MAIL: {itjeff, howard}@cityu.edu.hk

Abstract:

The Chinese characters evolved from pictograms and they are composed of strokes. A standard stroke sequence for each character is available in the dictionary. People introduced heuristic rules to specify the stroke order for easy memorization but it is very ambiguous to reconstruct the dictionary sequence according to the heuristic rules. In this paper, we combine the stroke extraction and stroke sequence reconstruction algorithms to reconstruct the strokes and their sequence from a Chinese character image. A well-known public Chinese character database (the HITPU database) is used as our input data. Performance evaluation shows the robustness of our proposed method and user evaluation shows that our proposed system helps users to create online Chinese character templates quickly and conveniently.

Keywords:

Stroke sequence estimation; stroke extraction; pattern classification; Chinese character; HITPU database

1. Introduction

The Chinese characters evolved from pictograms. The ancient Chinese used drawings to represent objects and some Chinese characters were evolved from these drawings. The more complicated the object, the more strokes were needed. Hence, stroke sequence and shape become crucial for proper Chinese handwriting. A generally accepted set of stroke sequence is defined in Chinese dictionaries. For easy memorization, people introduced heuristic rules to generalize the stroke order of most Chinese characters [1], for example, ordering the strokes from top to bottom and from left to right. However, the rules are limited in accuracy since large structural variations exist among the ten thousand Chinese characters.

Some researchers use the stroke sequence as a feature in online Chinese character recognition. Nakai et al. [2] use the Markov Chain to model the stroke sequence in their character recognizer. Chen et al. [3] define rules on the basis of traditional heuristic stroke sequence rules; also, construct a nearly unique stroke sequence for every character. However their generated stroke sequence is not always the same as dictionary sequence. Similarly Liu et al. [4] define sequence of stroke segments with radical hierarchy details. Hung et al. [5] use numbers to represent different stroke forms and hence apply the number sequence matching in their character recognizer.

Some researchers propose systems to facilitate students in practicing handwriting. A pen-based input device is adopted and the students' ability in mastering the dictionary stroke sequence is assessed [6][7][8]. Evaluation is done by comparing the student's input handwriting with the template character, which is the standard handwriting created by a skilled teacher. The system proposed by Tan et al. [6] is yet limited in flexibility as the features of each template character should be input to the database manually. The methods proposed by Tsang et al. [7] and Tang et al. [8] are comparatively more flexible in the way that they were able to compare the input character with the template in stroke-by-stroke manner. Hence, the input stroke sequence could be verified by stroke matching. On contrast, the existing systems required teachers to spend enormous effort on writing template characters one by one. It would be helpful if the templates could be generated automatically.

To breakthrough the above limitations observed, we are motivated to develop a method which could automatically generate online Chinese character templates with dictionary stroke sequence from offline character images. However, the stroke extraction and stroke sequence estimation methods are still needed since it is difficult to reconstruct the stroke sequence from character images and to extract the complete strokes from it.

Lin et al. [9] propose to use a video camera to capture user's hand movement when writing on ordinary paper, meanwhile, extract the strokes and writing sequence. Lau et al. [10] extract stroke segments in a cursive signature by breaking the intersection points of the thinned binary handwriting image. Lin et al. [11] extract stroke segments similarly and introduced a bi-direction graph method to connect the segments into complete strokes. We extend

1-4244-0973-X/07/$25.00 ©2007 IEEE

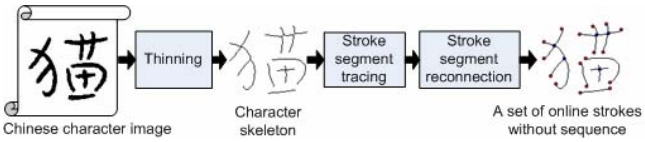
their work to extract strokes from a character image by removing the noise and spurs on the character skeleton.

Given a set of online strokes, researchers attempt to determine the dictionary sequence of the Chinese characters. Shimomura [12] considers minimization of the hand movement energy but it does not work well for characters with more than 6 strokes or complicated structure. Joe et al. [13] use a similar method to recognize Korean characters that have fewer strokes. Lau et al. [10] consider the writing direction of stroke segments but it is not consistent enough to define the stroke order in regular scripts.

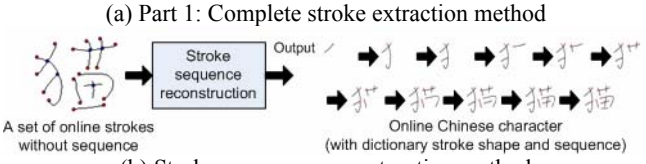
In this paper, we aim to reconstruct the strokes and their sequence from a Chinese character image. We combine the stroke extraction method proposed by Lin et al. [11] and the stroke sequence reconstruction method we proposed in [14]. We consider the HITPU database [15] as input data. Complete strokes are obtained by connecting stroke segments produced by regular stroke tracing. To reconstruct the stroke sequence, a classifier that defines the forward/backward stroke order is first trained. The classification result is then used to define a discrete state transition cost. A set of candidate stroke sequences with minimal total state transition cost is obtained. The candidate sequence that uses the least handwriting energy is chosen.

2. Proposed Method

In our proposed method, the input is a Chinese handwriting character image (offline character) and the output is an online Chinese character that both the stroke shape and sequence are conformed to the Chinese dictionary. Our method consists of two parts, the first part as shown in Figure 1(a), the character skeleton is first extracted by thinning. The stroke segments are traced out according to the feature points, the segments are then reconnected into complete strokes by extending the method proposed by Lin *et al.* [11]. In the second part as shown in Figure 1(b), the stroke sequence is reconstructed by training a set of positional features that define the microscopic order among the strokes.



(a) Part 1: Complete stroke extraction method

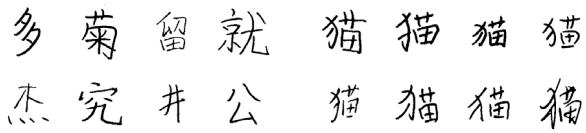


(b) Stroke sequence reconstruction method

Figure 1. Illustration of our proposed algorithm.

2.1. Our Input Dataset

The handwriting images in the HITPU database [15] are considered to be the input data, a well-known Chinese handwriting character database. It contains 751,000 samples of 3,755 different Chinese characters, which were written by 200 writers from Harbin Institute of Technology and Hong Kong Polytechnic University. Figure 2(a) shows some handwriting samples in the HITPU database. Figure 2(b) demonstrates the handwriting variations of a character "猫" ("Cat" in English) among different writers. The variation is meaningful in reflecting the robustness of our proposed method.



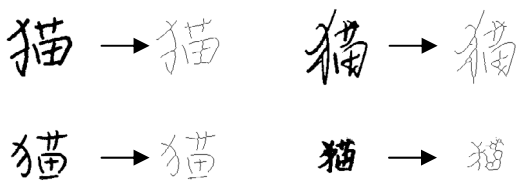
(a) Sample characters (b) Handwriting variations

Figure 2. Handwriting samples in HITPU database.

2.2. Stroke Extraction

In this part, a set of online strokes from a Chinese character image will be reconstructed. An online stroke is a series of data points in terms of (x, y) coordinates. The skeleton of the character image is first obtained by a thinning algorithm.

The handwriting samples in HITPU database vary in both writing style and quality. **Error! Not a valid bookmark self-reference.**(a) and **Error! Not a valid bookmark self-reference.**(b) show the thinning results of good quality samples and poor quality samples respectively. The strokes in good quality samples are clear with the correct structure, while the strokes in poor quality samples are always concatenated with each other. It is difficult to extract the complete strokes since the skeleton may form closed loops that are different from the strokes in actual handwriting. Therefore, in our experiment dataset, we only choose the handwriting samples that are in regular script, their qualities are better and will be sufficient to create handwriting templates.



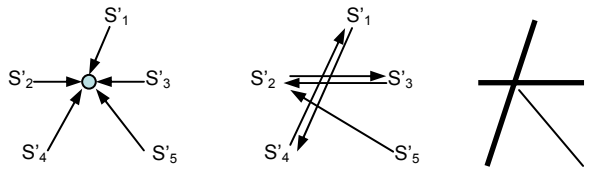
(a) Good quality samples (b) Poor quality samples

Figure 3. Thinning results of handwriting samples with different qualities.

The complete strokes are then extracted from the character skeleton by extending the method proposed by Lin *et al.* [11]. The stroke segments are first extracted by identifying the feature points on the skeleton. The feature points are end-points or the fork-points of the skeleton. The fork point is the point of intersection of the lines. Equation (1) shows the Rutoviz's crossing number $N_c(p)$ [16] that defines the adjacent connectivity of a pixel p. A pixel p is an end-point if $N_c(p)=1$. It is a fork-point if $N_c(p)>2$. The stroke segments by regular stroke tracing method, which traces out the lines in between the feature point,.

$$N_C(p) = \frac{1}{2}\sum_{i=1}^{8}\left|x_{i+1} - x_i\right| \qquad (1)$$

Next, the stroke segments that originally constitute the same complete stroke are re-connected. The stroke segments are merged at each fork pixel. Figure 4(a) shows a Chinese character "大" ("big" in English) that contains five stroke segments S'_i. The writing directions of these strokes towards the fork point are being considered. The stroke pairs with the most similar direction could form a link with each other, which is shown in Figure 4(b). In this case, the pairs (S'_1, S'_4) and (S'_2, S'_3) form bi-directional links. Figure 4(c) shows the final result that S'_1 connects with S'_4 and S'_2 connects with S'_3. The above steps repeats until all fork points are solved.



(a) The stroke segments towards a fork point

(b) The links of the stroke segment pairs

(c) Complete strokes are formed

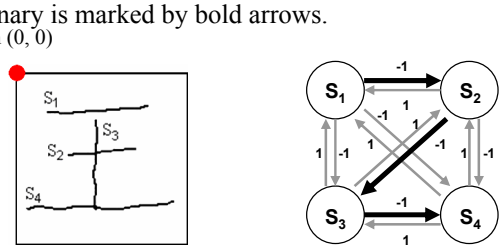
Figure 4. Illustration of stroke segment connection.

Some enhancements have been introduced when reconstructing the strokes. The noisy pixels are first filtered by removing pixel clusters with less than 4 pixels. The unwanted spurs at turning edges that are formed during thinning. They are pruned by successive thinning with dilation. It is also hard to ensure the stroke shape conform to the Chinese dictionary. For example, in dictionary the radical "□" is composed of three strokes: |□→□→□. The turning edge at the top-right corner may occasionally fail to re-connect due to handwriting variations.. It could be a future work to find a generic solution. At this moment, manual intervention is introduced to break/connect strokes at the corners when necessary. The writing direction of each reconstructed stroke is deduced by considering the absolute distances d from the end pixels on the stroke to the origin. The end pixel with a smaller value of d would be the starting point.

2.3. Stroke Sequence Reconstruction

A set of online strokes are extracted from a Chinese character image. Each stroke is labeled and the composing coordinates of which are ordered according to the writing direction and grouped under the same stroke label.

The dictionary sequence of the set of online strokes is then reconstructed. Figure 5(a) shows a sample Chinese handwriting character "王" ("King" in English) that contains 4 strokes labeled as S_1, S_2, S_3 and S_4. The top-left corner of the grid is the origin and each stroke is modeled as discrete state s_1, s_2, s_3 and s_4 respectively. Figure 5(b) shows the possible transitions among the states. A discrete transition cost is assigned to each transition that defined the microscopic order between any two states. A complete stroke sequence is just the minimum cost route that every state has been visited once without returning to the starting state. In the example case, we have 4 × 3 = 12 possible transitions. The ground truth route that obtained from dictionary is marked by bold arrows.



(a) The coordinate system (b) The dictionary sequence

Figure 5. Mapping between the strokes of the Chinese character "王" and the corresponding states.

The stroke sequence reconstruction method consists of three steps. For each possible transition, the microscopic order is classified. A pair of states could be in either forward of backward order. A forward order means the state s_i appears in front of s_j in the whole stroke sequence, while i and j are positive integers smaller than number of states n. The backward order is just the vice versa.

The state transition cost is the classifier derived from three offsets: the horizontal offset (H), vertical offset (V), and radial offset (R) as shown in Figure 6. The offset is the difference between the strokes of the corresponding pair of states. Figure 7 shows the feature points that characterize the shape of a stroke, which are the beginning point, the middle point and the ending point. To ensure the

classification rate symmetric, three symmetric point pairs (b_i, b_j), (m_i, m_j) and (e_i, e_j) are considered only. Hence, 3 × 3 = 9 different positional features are defined.

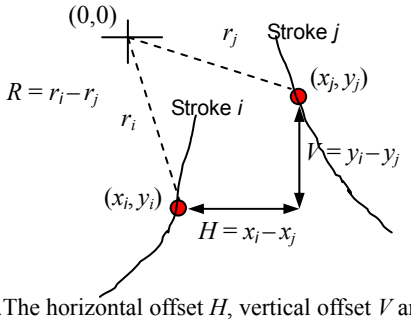


Figure 6.The horizontal offset H, vertical offset V and radial offset R.

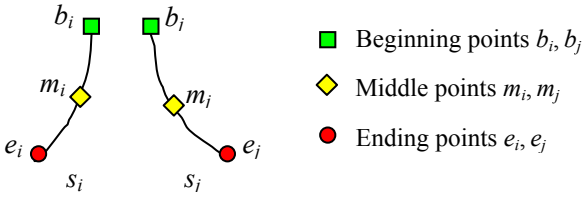


Figure 7. Beginning, middle and ending points for two example strokes s_i and s_j.

A set of 100 different characters from the HITPU database is selected as training data. The ground truth order of the stroke pairs are obtained directly from the dictionary. The classification rate of stroke order for each feature is obtained. The top six features are selected and combined linearly. The weight of each feature is calculated by multiplying the pseudo-inverse of the feature matrix M that contains all 6 selected features. Let v and w be the vectors containing the ground truth data and the weights of the features. The weight is hence:

$$w = M^{+}v \tag{2}$$

Let f_i and w_i be the feature value and its weight of each stroke pair respectively. The feature value f is hence:

$$f = \sum_{i=1}^{6} f_i w_i \tag{3}$$

The state transition cost is determined by the sign of the combined feature value f. The state transition cost becomes +1, -1 and 0 if f is positive, negative and zero respectively. A stroke sequence is the route with the minimum transition cost. It is equivalent to the Traveling Salesman Problem (TSP) [17]. Genetic Algorithm [18] is applied to enhance the efficiency. Figure 8 shows the possible results of a character contains 4 strokes. In this case, 4 possible sequences are obtained. The sequence with the least total cost becomes our result.

However, the discrete state transition cost results more than one minimal cost sequences, yet, the final decision is made by minimizing the total handwriting energy E_h, which models the hand movement when writing [12]. E_h is the sum of head-to-tail distances between adjacent strokes on the reconstructed sequence. The candidate sequence with minimal handwriting energy is the final result.

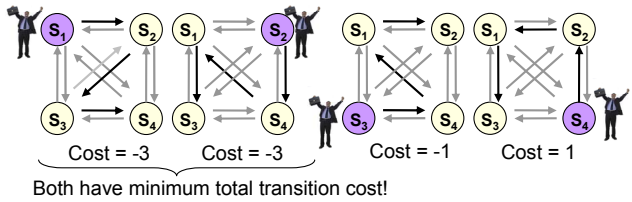


Figure 8 Minimal cost paths with different starting states

3. Experiments and Results

The 100 sets of different Chinese characters from HITPU database have been selected as input data. Table 1 shows these characters ordered by number of strokes. These characters have an average of 7.34 strokes per character that is close to the average stroke number of frequently used Chinese character in dictionary, which is 7.3 [19]. In each Chinese character, 15 best quality samples out of total 200 have been chosen. One sample is randomly selected as the training sample and the remainings used as testing samples. So, our dataset contains totally 1500 good samples with 100 training and 1400 testing samples respectively.

Table 1. Chinese characters in our dataset.

Number of stroke	Chinese character
2	丁九力
3	工久口亏弓川凡
4	斗公井巨亢毛孔丹
5	功句卡刊可立另令
6	多光臼扛扣夸劣芒忙
7	攻花戒究坎克快况困牢冷李利良吝伶卵
8	杰金京咎拘狙居咀刻肯空昆垃拉例
9	皇韭俊咯柯咳枯括俐洛茫
10	氦疾恐栗烈留浩浸
>10	竟菊啦琅淋琉曼猫景就腊湖督赫

The performance of our proposed method is compared with two existing methods: Shimomura [12] constructs stroke sequence with minimization of handwriting energy only and Lau *et al.* [10] considers the direction tendency between stroke segments. Figure 9 shows the percentage of reconstructed sequences that exactly the same as the dictionary stroke sequence. The reconstruction rate drops dramatically and they no longer reconstruct the exact dictionary sequence with more than 6 strokes. Our proposed method can reconstruct the exact dictionary sequence even

the character contains more than 10 strokes. The reconstruction rate of our method drops more gradual than existing works.

On the other hand, the rank distance is considered. It measures the similarity between the reconstructed sequence and the dictionary sequence. Our previous work [14] used the Kendall distance [20] as the rank distance that counts the minimum swaps of the stroke pairs in the reconstructed sequence until the dictionary sequence is reached. Suppose the reconstructed sequence is {1, 4, 2, 3} and the dictionary sequence is {1, 2, 3, 4}, three swaps are required. Our rank distance is enhanced by grouping the correct stroke sub-sequences and counts the minimum swaps of groups. From which, 3 groups ({1}, {4} and {2, 3}) are formed in the reconstructed sequence. Only the groups {4} and {2, 3} have to be swapped and hence the rank distance become 1. Figure 10 compares the rank distances among all stated methods. Our proposed method yields smaller rank distances. It shows that our reconstructed stroke sequence is the most conform to the dictionary sequence.

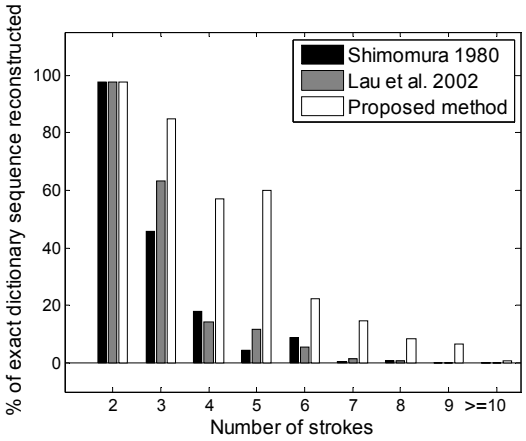


Figure 9. Comparison of the rate of fully reconstructed stroke sequence.

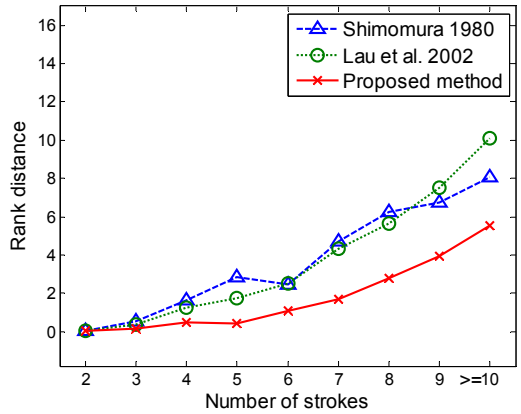


Figure 10. Comparison of the rank distances.

Online Chinese handwriting character is meaningful on education and research usages. An online Chinese character database can be created by simply writing through a pen-based input device but it takes time and energy to write ten thousand characters in a neat manner. Our proposed method is more convenient in creating character templates. The Chinese character images could be converted into online format and their dictionary stroke sequences could also be reconstructed. Although a few strokes may still out of order, these strokes could be swapped manually. A user test is performed to compare our proposed method with simply writing.

We have invited 7 university students to prepare online Chinese character templates with two methods. First, they have to create 30 Chinese characters as shown in Figure 11(a) by our proposed method, with a simple tool that helps them to swap strokes when necessary. As control experiment, they are then told to create other 30 Chinesecharacters as shown in Figure 11(b) by simply writing through a pen-based input device. They have to input the characters neatly and clearly, otherwise, they have to re-write again. Figure 11 indicates the finishing time of each task. On an average, people spend nearly a double of time on simply the writing task. The result would be more significant for people to build a large database with ten thousand Chinese characters.

Experiment	Characters	Average number of stroke	Finishing time (s)	
			Mean	Standard Deviation
(a) Proposed method	九枯括狙句猫花京啦刻烈牢氨利扛肯恐咀亏凡曼督另拘坎杰令咳戒腊	7.87	426.1	75
(b) Control method	川丁弓浩赫湖皇竟韭咎俊空扣快困垃拉冷栗俐立良劣淋吝琉洛茫忙毛	7.83	761.7	208

Figure 11. User test results.

4. Conclusion and Future Work

We propose a method to reconstruct strokes and stroke sequence from Chinese handwriting character images. The complete strokes are reconstructed by connecting the stroke segments obtained from the character skeleton. To reconstruct the stroke sequence that conforms to the dictionary sequence, the strokes are modeled as discrete states on a graph. The state transition cost is derived from the positional features. The candidate stroke sequences are solved by Traveling Salesman Problem. Through minimizing the total handwriting energy, the final sequence

is obtained. The experiment results have proven that our method could estimate stroke sequence more conform to dictionary among existing methods. In this sense, the users could create Chinese character templates more efficiently with our method than simply writing through a pen-based device.

In future work, more features related to stroke order will be explored. The effects of different weighting for different stroke types and radical decomposition will be studied. Our method is also suggested to be applied in other Asian character sets like Japanese Hiragana/Katakana and Korean Hangul since they are also stroke based characters.

Acknowledgements

The work described in this paper is fully supported by a grant from City University of Hong Kong (Project No. 7001711).

References

[1] The structures and stroke sequence rules of Hanzi (In Chinese), The Commercial Press (HK) Ltd. [Online]. Available: http://www.cp-edu.com/TW/CIKU/free_html/fl_hzjglx.asp. [Accessed: Mar. 21, 2007].

[2] Mitsuru Nakai, Naoto Akira, Hiroshi Shimodaira and Shigeki Sagayama, "Substroke Approach to HMM-based On-line Kanji Handwriting Recognition", 6th Intl. Conf. on Document Analysis and Recognition, pp. 491–495, Sep. 2001.

[3] Zen Chen, Chi-Wei Lee and Rei-Hen Cheng, "Handwritten Chinese character analysis and preclassification using stroke structural sequence", Proc. of the 13th Intl. Conf. on Pattern Recognition, vol. 3, pp. 89–93, 1996.

[4] Ying-Jian Liu, Li-Qin Zhang and Ju-Wei Tai, "A new approach to on-line handwritten Chinese character recognition", Proc. of the 2nd Intl. Conf. on Document Analysis and Recognition, pp. 192–195, 1993.

[5] Kwok-Wah Hung, Wing-Nin Leung and Yau-Chuen Lai, "Boxing code for stroke-order free handprinted Chinese character recognition", IEEE Intl. Conf. on Systems, Man, and Cybernetics, vol. 4, pp. 2721–2724, 8–11 October 2000.

[6] Chwee Keng Tan, "An algorithm for online strokes verification of Chinese characters using discrete features", 8th Intl. Workshop on Frontiers in Handwriting Recognition, pp. 339–344, 2002.

[7] Kerry Tsang and Howard Leung, "Teaching Stroke Order for Chinese Characters by Using Minimal Feedback", Intl. Conf. on Web-based Learning (ICWL 2005), Hong Kong, August 2005.

[8] Kai-Tai Tang, Ka-Ki Li and Howard Leung, "A Web-based Chinese Handwriting Education System with Automatic Feedback and Analysis", 5th Intl. Conf. on Web-based Learning, Malaysia, July 2006.

[9] Feng Lin and Xiaoou Tang, "Dynamic stroke information analysis for video-based handwritten Chinese character recognition", Proc. of the 9th IEEE Intl. Conf. on Computer Vision, vol. 1, pp. 695–700, 2003.

[10] Kai-Kwong Lau, Pong-Chi Yuen and Yuan Yan Tang, "Universal Writing Model for Recovery of Writing Sequence of Static Handwriting Images", Intl. Journal of Pattern Recognition and Artificial Intelligence, vol. 19, no.5, pp. 1–27, 2005.

[11] Feng Lin and Xiaoou Tang, "Off-line handwritten Chinese character stroke extraction", Proc. of the 16th Intl. Conf. on Pattern Recognition, vol. 3, pp. 249–252, 2002.

[12] Takeshi Shimomura, "Science of the stroke sequence of Kanji", 8th Intl. Conf. on Computational Linguistics, pp. 270–273, 1980.

[13] Moon Jeung Joe, Huen Joo Lee, "A combined method on the handwritten character recognition", Proc. of the 3rd Intl. Conf. on Document Analysis and Recognition, vol. 1, pp. 112–115, August 1995.

[14] Kai-Tai Tang and Howard Leung, Reconstructing the Correct Writing Sequence from a Set of Chinese Character Strokes, Proc. of the ICCPOL 2006 (LNCS 4285), Springer, 2006, pp. 333–344.

[15] Daming Shi and Bob Damper, HITPU database. [Online]. Available: http://www.ntu.edu.sg/home/ asdmshi/hitpu.html. [Accessed: Mar. 21, 2007].

[16] Louisa Lam, Seong-Whan Lee and Ching Y. Suen, "Thinning methodologies – a comprehensive survey", IEEE Trans. on PAMI, vol. 14, no. 9, pp. 869–885, 1992.

[17] D. S. Johnson and L. A. McGeoch, The Traveling Salesman Problem: A Case Study in Local Optimization, Local Search in Combinatorial Optimization, E. H. L. Aarts and J.K. Lenstra (ed), John Wiley and Sons Ltd, 1997, pp 215–310.

[18] J. Grefenstette, R. Gopal, R. Rosmaita, and D. Gucht, "Genetic algorithms for the traveling salesman problem", Proc. of the 2nd Intl. Conf. on Genetic Algorithms, Lawrence Eribaum Associates, Mahwah, NJ, 1985.

[19] Shi-Zhao Zhang, The statistics of Chinese character strokes (In Chinese). [Online]. Available: http:// www.chancezoo.net/hz/hzbhtjtx.htm. [Accessed: Mar. 21, 2007]

[20] Kardi Teknomo, "Similarity Measurement". [Online]. Available: http://people.revoledu.com/kardi/tutorial/ Similarity/. [Accessed: Mar. 21, 2007].