Proceedings of the IEEE
International Conference on Automation and Logistics
August 18 - 21, 2007, Jinan, China

Segmentation of On-line Cursive Handwritten Chinese Word Based on Stroke Speed Feature and Stroke Vector Feature*

GUO Rui, JIN LianWen+

School of Electronics and Information Engineering
South China University of Technology
Guangzhou, Guangdong Province, China
+E-mail: eelwjin@scut.edu.cn

***Abstract* - On-line handwritten Chinese word recognition has recently become an important research topic in the filed of computer vision. However, the segmentation of cursive Chinese word is still an unsolved problem. In this paper, two new features, Stroke Speed Feature and Stroke Vector Feature, are proposed for the segmentation of on-line handwritten Chinese word. Analysis and experiments show that both of the features are easy to implement, with low computation complexity and encouraging correct segmentation accuracy. Furthermore, the Stroke Vector Feature outperforms traditional histogram method and we found it is especially suitable for the segmentation of cursive handwritten word where two characters touch each other or overlap.**

Index Terms - On-line Chinese character segmentation, Word recognition, Stroke Vector Feature, Stroke Speed Feature.

I. INTRODUCTION

Character segmentation has been a challenging problem in unconstrained handwritten Chinese character recognition for a long time. As the recognition system is currently being developed to recognize cursive continuous characters (e.g. word recognition), it is necessary to find an effective method to separate the characters automatically. Actually, character segmentation has become an essential part of on-line handwritten Chinese word recognition [1][2].

Many approaches have been reported for the segmentation of unconstrained handwritten Chinese characters. Such as histogram projection, connected components, stroke bounding boxes, recognition-based segmentation and holistic [3][4][9]. However, most of them are focused on the off-line handwritten Chinese characters segmentation and few studies are related to the on-line cursive Chinese characters segmentation. For off-line handwritten Chinese character segmentation, many methods use the prior knowledge about Chinese character structure [5][7][10]. The segmentation problem that characters may be written to touch each other or to overlap with each other is still not an easy one [4][9]. For on-line handwritten Chinese character segmentation , there are only a few studies reported and the methods proposed in these papers often demand high computation complexity[1][8]. Furthermore, little attention is paid on the use of the on-line information of the Chinese character strokes for segmentation.

In this paper, we focus on the on-line unconstrained handwritten Chinese word segmentation and propose a new point of view for the on-line handwritten Chinese character segmentation. We adopt the sequence information of strokes to segment the unconstrained handwritten Chinese word samples. According to the investigation of a large number of unconstrained handwritten Chinese word samples, two segmentation features, Stroke Speed Feature (SSF) and Stroke Vector Feature (SVF), are proposed. Both of them are with low computation complexity and easy to implement, especially for the Stroke Vector Feature. Compared with the conventional histogram projection method, Stroke Vector Feature has a similar computation complexity but a better performance. We found that these segmentation features are useful for the coarse segmentation stage. Based on the observation of the structural characteristics of Chinese handwritings and experimental results, we tentatively put forward a conclusion that by using the SVF we can detect the last stroke of the previous charachter with a good accuracy. It may be the reason for the good performance of Stroke Vector Feature when two characters touch each other or overlap.

The remaining parts of the paper are organized as follows. In Section II**,** we describe the two new segmentation features respectively. Experimental results and discussion are given in Section III and we conclude the paper in Section IV.

II. SEGMENTATION BASED ON STROKE SPEED FEATURE AND STROKE VECTOR FEATURE

For on-line Chinese character recognition, the segmentation is an important pre-processing because correct recognition of characters relies on correct segmentation of characters. In order to achieve good performance in segmenting unconstrained handwritten Chinese word, the segmentation approach should take account of some special properties of Chinese characters. We studied a large number of unconstrained handwritten Chinese word samples and finally found out two new segmentation features that not only can be computed easily but also have good correct segmentation rate, which will be described as follows.

* This paper is partially sponsored by the following foundation: New Century Excellent Talent Program of MOE (No.NCET-05-0736); The University Fund of Microsoft Research Asia (No.FY07-RES-THEME-58)

A. Stroke Vector Feature (SVF)

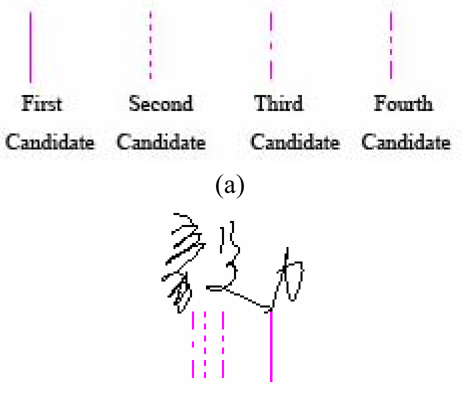
A handwritten word with N strokes is denoted as $W_N = \{S_1, S_2, S_3, ...S_N\}$, and each stroke S_i consists of a series of points: $S_i = \{P_1 P_2 P_3 ...P_t\}$. The proposed Stroke Vector Feature (SVF) between stroke S_i and S_{i+1} is defined by:

$$Vec_{i,i+1} = Dis(P_i^b, P_{i+1}^b) \qquad (1)$$

Where the P_i^b represents the first point of the ith stroke S_i, and the Dis function is used to calculate the city-block distance between the corresponding two points. A feature value sequence can be then obtained as:

$$V = \{Vec_{1,2}, Vec_{2,3}, Vec_{3,4}, ...Vec_{N-1,N}\} \qquad (2)$$

Several segmentation candidate strokes are chosen by sorting the sequence in descending order according to the feature value. The first one from the sorted sequence is denoted as the first candidate stroke and etc. The rightmost points of the candidate strokes are then extracted and taken as candidate segmentation points through each of which a segmentation path may be formed, as shown in Fig. 1(b). For a clear view of different candidate segmentation points, we use different line styles to represent the candidate sequence, as shown in Fig.1(a).



(a)

(b)

Fig. 1 (a) Line styles (b) Handwritten word "毅力" and each candidate segmentation path, this word consists of 5 strokes.

As it is known, an on-line handwritten Chinese character is an ideograph and composed of sequential strokes. Based on the observation of the structural characteristics of Chinese handwritings and experimental results, the value between two characters' strokes is normally greater than the values between each character's internal strokes. Therefore, we can segment the word samples effectively by using this property via SVF and this approach is also robust for the segmentation task that characters may be written to touch each other or to overlap with each other.

B. Stroke Speed Feature (SSF)

We denote the elapsed time between the first points of the stroke S_i and S_{i+1} as T_i, for a word $W_N = \{S_1, S_2, S_3, ...S_N\}$ and $S_i = \{P_1 P_2 P_3 ...P_t\}$, then the Stroke Speed Feature (SSF) sequence is defined as follows:

$$L_{i,i+1} = Dis(P_1, P_2) + Dis(P_2, P_3) + ... + Dis(P_t, P_{i+1}^b) \qquad (3)$$

$$SP_i = L_{i,i+1} / T_i \qquad (4)$$

$$aSP_i = (SP_{i+1} - SP_i) / T_i \qquad (5)$$

We extract aSP_i from the feature value sequence: $aSP = \{aSP_1, aSP_2, ...aSP_{N-2}\}$ only if aSP_i <0 and sort them in ascending order, the first one of the sorted sequence is denoted as the first candidate stroke. Similarly, the rightmost points of the candidate strokes are then extracted and taken as candidate segmentation points, as shown in Fig. 2.

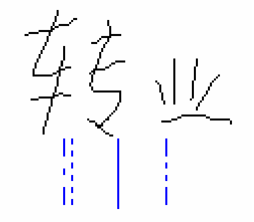


Fig. 2 Handwritten word "转业" and each candidate segmentation path, this word consists of 13 strokes.

III. Experimental Results and Discussion

15 sets of uncontrained cursive handwritten Chinese word samples were collected by us using PDA(personal digital assistant) and Tablet PC. Each set consists of 14822 two-character words, 1195 three-character words and 1349 four-character words. The writers of the samples include postgraduates, undergraduate students and workers.

The data used in the experiments are part of the whole handwritten word database. From the two-character, three-character, and four-character word samples of each set, we randomly chose 500 samples respectively to test the two features. Some examples of the data are illustrated in Fig. 3.

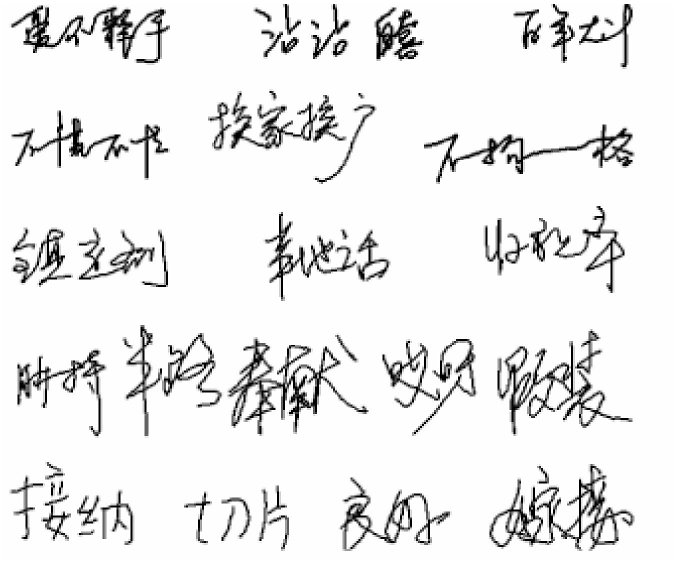


Fig. 3 Examples of handwritten words we used

A. *Experimental Results*

We use the correct segmentation rate (CSR), i.e. the ratio of correctly segmented word samples among all reference segmented word samples, to evaluate the performance of Stroke Vector Feature and Stroke Speed Feature. Table I gives the CSR of first candidate segmentation results.

TABLE I
SEGMENTATION RESULTS USING THE PROPOSED TWO FEATURES

CSR (%)	Two-char	Three-char	Four-char
SVF	64.73%	43.08%	31.55%
SSF	39.10%	19.55%	9.66%

From Table I, it can be seen that the correct segmentation rate of SVF is higher than SSF. For two-character word, the best result achieved is 64.73%. Although this rate is not high enough, but considering the tough segmentation task we faced (cursive handwritten Chinese word that are written freely in an unconstraint manner), the results we got using the novel features is still encouraging.

As many popular segmentation algorithms contain a coarse segmentation stage, it is necessary to find some approach to generate several candidate segmentation points effectively. Thus, using the same test set as Table I, we carried out some experiments to evaluate the performace of these two features when increasing the number of candidates. The results are shown in Table II. The leftmost column denotes the number of candidates based on the first candidate segmentation path. It can be seen that SVF is superior to SSF. When we supply 3 more candidates for segmentation, the CSRs of SVF are improved to 85.06%, 79.55% and 74.56% for two-character, three-character and four-character word respectively. If we increase the number of candidates, the CSR will be higher and the contextual and recognition information may be used to verify the candidate segmentation paths. For SSF, the correct rate sometimes doesn't increase but even decrease, this is due to the fact that a great number of samples couldn't obtain enough number of candidates for $aSP_i < 0$ because many of them are written too cursive.

TABLE II
SEGMENTATION RESULTS WHEN INCREASING THE NUMBER OF CANDIDATES

	Stroke Vector Feature			Stroke Speed Feature		
CSR (%)	Two-char	Three-char	Four-char	Two-char	Three-char	Four-char
+1	78.50%	61.52%	50.53%	51.60%	34.65%	24.24%
+2	84.75%	73.72%	64.83%	49.02%	41.44%	37.01%
+3	85.06%	79.55%	74.56%	38.15%	40.34%	39.34%

In another experiment, we randomly chose 2000 two-character word samples from the whole database to compare the performance of our method against histogram projection method, which is a frequently used basic method in handwritten characters segmentation [2][4]. Table III gives the results.

TABLE III
COMPARISON WITH TRADITIONAL HISTOGRAM PROJECTION

CSR (%)	Histogram Projection	SVF	SSF
Top 1	56.79%	67.36%	44.43%
+1	66.29%	81.24%	56.79%
+2	73.86%	86.79%	46.79%
+3	78.21%	87.84%	34.50%

From Table III, it can be seen that, the proposed method using SVF has a significant improvement of 10.57% CSR than histogram projection method for the top 1 candidate segmentation result. For the situations of more than one additional segmentation candidates, SVF still outperform the histogram method, showing that the segmentation method using SVF is effective. However, the performance of SSF is not good enough, but it provides some alternative segmentation information which might be integrated with SVF to produce much better performance, which worth our further study.

B. *Discussion on Stroke Vector Feature*

In unconstrained cursive handwritten Chinese characters, characters may be written to touch each other or to overlap with each other, therefore the segmentation problem is a tough task.

After observing a large number of handwritten Chinese word samples segmented with SVF, we found that this feature is good for the previous segmentation problem. With a small computational cost, the SVF can segment the word whose characters touch each other or overlap with good accuracy. Based on the analysis of the structural characteristics of Chinese handwritings and experimental results, it is found that last stroke of the previous charachter in a Chinese word (usually this is a valid segmentation point) can be detected using SVF with a high accuracy.

If the last stroke of the previous charachter is detected correctly, then, with the stroke sequence information of the word, the previous character can be recovered from the word and the characters can be segemented perfectly. It is worth to noting that to segment character using SVF is based on the hypothesis that there is pen up when one character is written completely. In most situations, people writing habits is in according with this hypothesis. Some detail examples with forward recovering results (FRR) at each candidate segmentation points are shown in Fig. 4 , Fig. 5 and Fig. 6. It can be seen that although the characters in these words touch or overlap heavily, we can segment the correct character using SVF successfully.

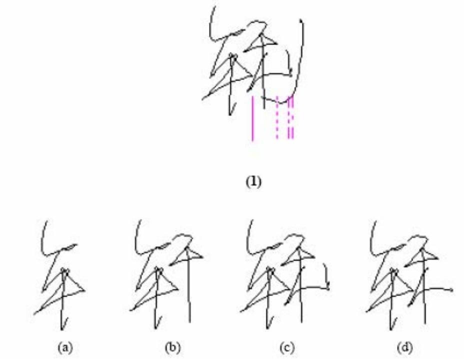


Fig. 4 (1) The word”牟利” and candidate segmentation paths (a) FRR at the top 1 candidate point (b) FRR at the second candidate point (c) FRR at the third candidate point (d) FRR at the fourth candidate point.

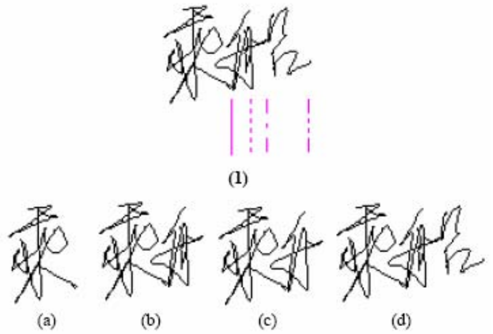


Fig. 5 (1) The word”乘船” and candidate segmentation paths (a) FRR at the top 1 candidate point (b) FRR at the second candidate point (c) FRR at the third candidate point (d) FRR at the fourth candidate point.

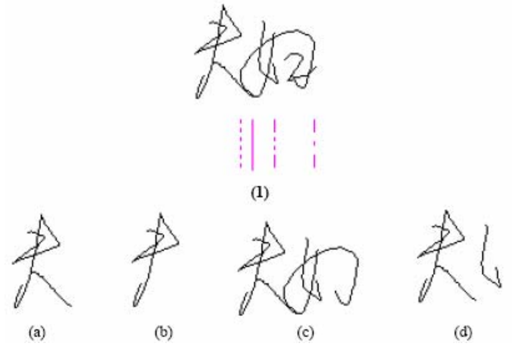


Fig. 6 (1) The word”夫妇”and candidate segmentation paths (a) FRR at the top 1 candidate point (b) FRR at the second candidate point (c) FRR at the third candidate point (d) FRR at the fourth candidate point.

IV. Conclusion

In this paper, two novel features that make use of the on-line writing information, Stroke Vector Feature (SVF) and Stroke Speed Feature (SSF), have been proposed for on-line handwritten Chinese word segmentation. Both of them are based on the stroke sequence information, which is obviously different from the conventional segmentation methods. Experimental results show that the SVF has a better performance than traditional histogram method. SVF is found especially effective for the segmentaion of word in which two characters touch each other or overlap. Moreover, these two segmentation features have low computation complexity and can be used at the coarse segmentation stage of many popular segmentation algorithms. Though the performance of SSF is not good enough, but it provides some alternative segmentation information which might be integrated with SVF to produce much better performance, which worth our further study. However, to solve the segmentation completely and obtain a high enough performance still need improving both of the features and may need more elaborate processing such as employing a classifier to guide segmentation or using large vocabulary lexicon to verify and correct the segmentation results.

References

[1] Zhengbin Yao, Xiaoqing Ding and Changsong Liu, “On-line handwritten Chinese word recognition based on lexicon,” *Proc. 18th Int. Conf. Pattern Recognition*, vol. 2, pp. 320-323, August 2006.

[2] Xiaoqing Ding, “Chinese Character Recognition: A Review,” *Acta Electronica Sinica*, vol. 30, no. 9, pp. 1364-1368, 2002.

[3] R.G. Casey, and E. Lecolinet, “A Survey of Method and Strategies in Character Segmentation,” *IEEE Trans. PAMI*, vol. 18, no.7, pp. 690-706, 1996.

[4] Jie Shao, Yu Cheng, “A Survey of Methods in Handwritten Chinese Character Segmentation,” *Computer Technology and Development*, vol. 16, no. 6, pp. 184-186, June 2006.

[5] Lin Yu Tseng, Rung Ching Chen, “A New Method for Segmenting Handwritten Chinese Characters,” *Proc. 4th Int. Conf. Document Analysis and Recognition*, vol. 2, pp. 568-571, 1997.

[6] Shourui Tian,Gengjian Ma, Yali Wang, Shaowei Xia, “Unconstrained Handwritten Chinese String Recognition System for the Amount on Bank Checks,” *Journal Tsinghua Univ(Sci&Tech)*, vol. 42, no. 9, pp. 1228-1232, 2002.

[7] Yi-Hong Tseng, Hsi-Jian Lee, “Recognition-based handwritten Chinese character segmentation using a probabilistic Viterbi algorithm,” *Pattern Recognition Letters*, vol. 20, no. 8, pp. 791-806, August 1999.

[8] Xue Gao, Pierre Michel Lallican, Christian Viard-Gaudin, “A Two-stage Online Handwritten Chinese Character Segmentation Algorithm Based on Dynamic Programming,” *Proc. 8th Int. Conf. Document Analysis and Recognition*, pp. 735-739, 2005.

[9] Yanyu Gao, Yang yang, “Survey of Unconstrained Handwritten Chinese Character Segmentation,” *Computer Engineering*, vol. 30, no. 5, pp. 144 - 146, 2004.

[10]Shuyan Zhao, Zheru Chi, Pengfei Shi and Qing Wang, “Handwritten Chinese character segmentation using a two-stage approach,” *Proc. 6th Int. Conf. Document Analysis and Recognition*, pp. 179-183, September 2001.

[11]C.-L. Liu, S. Jaeger, and M. Nakagawa, “Online recognition of Chinese characters: the state-of-the-art,” *IEEE Trans. PAMI*, vol. 26, no. 2, pp. 198-213, February 2004.

[12]J. Cai and Z. Liu, “Off-Line Unconstrained Handwritten Word Recognition,” *Australian and New Zealand Conf. on Intelligent Information Systems*, pp. 199-202, November 1996.

[13]Yan Jiang, Xiaoqing Ding, Qiang Fu, and Zheng Ren, “Context Driven Chinese String Segmentation and Recognition,” *SSPR&SPR 2006*, pp. 127-135, 2006.