Available online at www.sciencedirect.com

SciVerse ScienceDirect

AASRI Procedia 1 (2012) 14 – 20

AASRI Procedia

www.elsevier.com/locate/procedia

2012 AASRI Conference on Computational Intelligence and Bioinformatics

A New Approach Based on Ncut Clustering Algorithm for Signature Segmentation

Jun Tan[a,*], Wen-Xian Wang[b], Min-Shui Feng [c] ,and Xiao-Xiong Zuo[b]

[a]*School of Information Science and Technology, Sun Yat-sen UniversityGuangzhou, P. R. China*
[b]*Public Security of Yuexiu, Guangzhou, P. R. China*
[c]*Public Security of Guangdong Province, Guangzhou, P. R. China.*

Abstract

In signature recognition, it is a significant step to segment a text line into characters. And the unsupervised clustering is a commonly used remedy for this task. However, due to the strong overlapping and touch among characters, the separation boundaries between two characters are usually nonlinear, which leads to the failure of the widely used clustering methods. To tackle this problem, this paper proposes a new signature segmentation method, which can effectively segment nonlinearly separable characters. In the proposed approach, we first segment the entire signature into strokes, the similarity matrix of which is computed according to stroke gravities. Then, the nonlinear clustering method termed Normalized cut (Ncut) is performed on this similarity matrix to obtain cluster labels for these strokes, through which the strokes are combined into characters. Experiments on four databases are conducted to demonstrate the effectiveness of the proposed method.

Selection and/or peer review under responsibility of American Applied Science Research Institute

Keywords: signature segmentation, overlapping, touch,nonlinearly separable, nonlinear clustering

1. Introduction

Signature recognition plays an important role in the fields such as forensic [1], historic document analysis [2], archaeology [3],etc. Segmentation is an indispensable step in signature recognition. However, off-line segmentation is not an easy task for two reasons. First, the number of characters is quite huge (e.g., there are

* Corresponding author. Tel.: 8620-8411-1188; fax: 8620-8411-1161.
E-mail address: mcstj@mail.sysu.edu.cn.

doi:10.1016/j.aasri.2012.06.004

more than 9000 frequently used Chinese characters). Second, there exist various writing styles, which makes both text line and single character segmentations very difficult. This paper mainly focuses on segmenting a text line comprising overlapping characters of arbitrary style into single characters.

Some methods have been proposed recently. Tseng and Lee [4] presented a recognition-based signature segmentation method for handwritten Chinese, This method firstly initializes several possible nonlinear segmentation paths by using a probabilistic Viterbi algorithm. A segmentation graph is constructed using these candidate paths with the shortest path finally detected as the segmentation path. In [5], the strokes are used to build stroke bounding boxes (SBB), which are merged by some knowledge-based merging operations, and a dynamic programming approach is used to find the best segmentation boundaries.

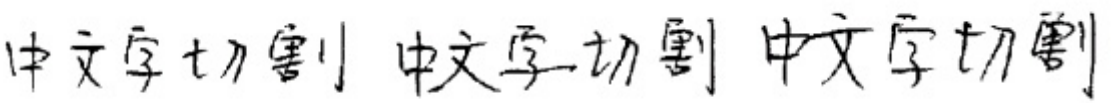
In [6], Yamaguchi et al. proposed a segmentation method for touching Japanese handwritten signature, which decreases over-segmentation by utilizing connecting condition of lines at the touching point. Wang et al. [7] proposed a method for Chinese signature segmentation based on connected components algorithm (CCA) template. This method first finds a black (foreground) pixel as the center of 3 × 3 template, from which the search of neighbor foreground pixels begins, until no new neighbor foreground pixel can be found.

In this paper, we propose a new signature segmentation algorithm based on nonlinear clustering. The proposed approach first factorizes each text line into strokes, the similarity matrix of which is computed. Then the nonlinear Normalized cut (Ncut) [8] clustering method is applied on this similarity matrix to obtain cluster labels for these strokes. The segmented characters can be finally achieved by combining strokes according to their cluster labels and positions.

The remainder of the paper is organized as follows. Section 2 introduces the preliminary knowledge of this work. In section 3, we describe in detail the proposed method. Experimental results are reported in section 4. We conclude our paper in section 5.

2. Preliminary Knowledge

This section introduces the preliminary knowledge of this work, including the concepts of signature segmentation, the nonlinearly separable problem suffered by handwritten signature segmentation and the nonlinear clustering methods such as Normalized cut (Ncut) [8].



(a) Type one (b) Type two (c) Type three

Fig. 1. Three types of character segmentation.

Due to the various writing styles, which often leads to touch and overlapping between two characters, According to the distance and separability between two characters, character segmentation is classified into the following three categories.

– Type one: Characters are isolated from each other, as shown in Fig. 1(a). The radicals forming different characters can be easily separated according to which character they belong to.

– Type two: Characters do not touch with each other, but are not linearly separable by vertical straight lines due to overlapping, as shown in Fig. 1(b).

– Type three: Characters touch with each other, as shown in Fig. 1(c).

Among the above three types, type two and type three are more challenging than type one. In this paper, we mainly focus on type two. The major difficulty of type two is that the separation gap between two characters are nonlinear, which cannot be achieved by a vertical straight line. Previous methods [5–7] are either on

segmenting characters that are linearly separable using a vertical straight line or on directly finding nonlinear segmentation paths. As shown in Fig. 2, directly finding nonlinear segmentation paths is a NP hard task in that there exist plenty of candidate segmentation paths without mathematical model.

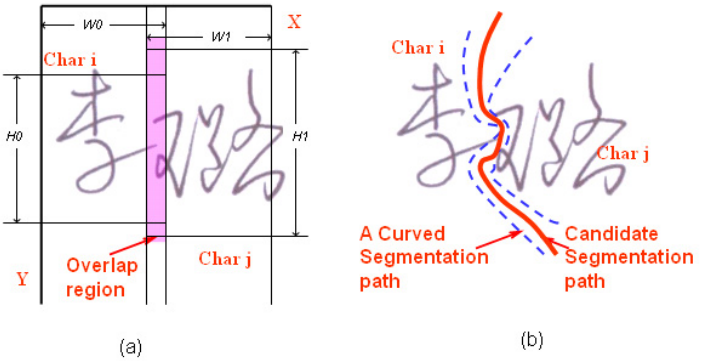


Fig. 2. The overlap region and candidate segmentation paths.

One efficient remedy for selecting correct segmentation paths from candidates is clustering. The basic idea is to over-segment one text line into many strokes, forming a stroke set. Then a clustering method is performed on this stroke set to generate cluster labels for these strokes. Finally, these strokes are combined as characters according to their cluster labels. For instance, Wang and Fan [9] proposed to use the classic k-means algorithm [10] to cluster strokes. However,it can only deal with linearly separable characters due to the linear limitation of k-means.

3. The Proposed Method

One method for tackling nonlinear separability is to use nonlinear clustering methods such as Normalized cut (Ncut) [8] and Conscience On-line Learning (COLL) [11]. In this paper, we utilize the well-known Ncut method.Figure 3 summarizes the overall flowchart of the proposed signature segmentation method, with some of the main procedures described as follows.

The scanned images of text are often gray-scale.We should first convert them into binary images. Then we utilize the thinning method proposed in [12] to obtain the skeleton of unitary thinness.

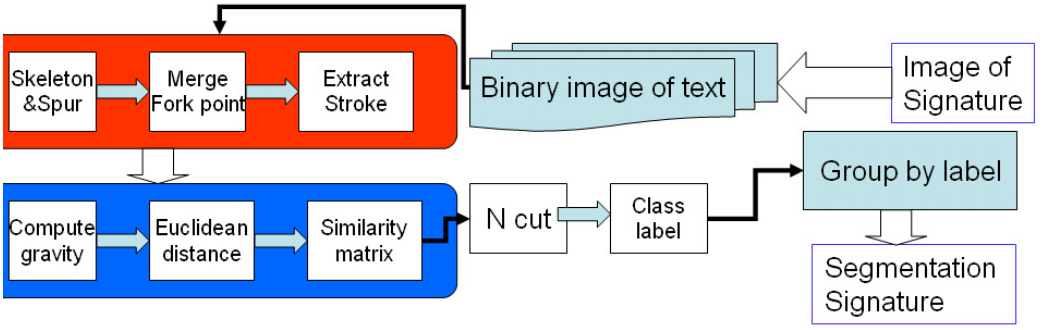


Fig. 3. Flowchart of the proposed signature segmentation method.

By scanning the preprocessed image, all the strokes can be extracted. Let $\{S_1, S_2, \ldots, S_n\}$ be n strokes

obtained from an image C. We first compute the gravity $\left\langle G_x^k, G_y^k \right\rangle$ of each stroke S_k,

$$G_x^k = (\sum_{i=1}^{M} i \times P_x(i)) / (\sum_{i=1}^{M} P_x(i)) \quad (1)$$

$$G_y^k = (\sum_{i=1}^{N} i \times P_y(i)) / (\sum_{i=1}^{N} P_y(i)) \quad (2)$$

where M and N are the width and height of S_k respectively.

The Euclidean distance between two strokes S_i and S_j are then defined as

$$f(S_i, S_j) = \left\| S_i - S_j \right\| = \sqrt{(G_x^i - G_x^j)^2 + (G_y^i - G_y^j)^2} \quad (3)$$

The Euclidean distance $f(S_i, S_j)$ measures the spatial relation between S_i and S_j . Accordingly, a similarity matrix $W = [w_{ij}]_{n \times n}$ is computed [8]

$$w_{ij} = \begin{cases} e^{\frac{-f(s_i, s_j)^2}{\sigma_c}} & \text{if} \quad f(S_i, S_j) < r \\ 0 & \text{otherwise,} \end{cases} \quad (4)$$

where σ_c is the variance of all the distances between strokes contained in image C and r is a threshold, say, the mean value of all distances.

The nonlinear clustering algorithm Ncut is performed on this similarity matrix W to obtain the cluster labels $\{L_1, L_2, \ldots, L_n\}$ for the n strokes. These strokes are combined to form characters according to their cluster labels and positions.

4. Experimental Results

4.1. Testing Databases

Four widely tested databases are used in our experiments. They are IAM [13],KAIST Hanja1 [14], HCL2000 [15] and SYSU [11]. The IAM database includes 1066 forms produced by 350 different writers and there are a total of 82227 word instances out of a vocabulary of 10841 words [13]. The KAIST Hanja1 database was collected by the Korea Advanced Institute of Science and Technology. It contains 783 frequently used Chinese characters, each character consisting of 200 samples written by 200 writers respectively [14]. The HCL2000 database The database contains 3755 frequently used simplified Chinese characters written by 1000 different writers [15].

(a) IAM [13] (b) KAIST Hanja1 [14] (c) HCL2000 [15] (d) SYSU [11]

Fig. 4. Some samples of the four databases.

SYSU database which is generated and collected by ourselves as follows. 245 volunteers were asked to sign

his (or her) name and one of the others' names twice. And a correction of 60025 signatures are obtained, which is named SYSU signature database [11]. Figure 4 shows some samples of databases.

4.2. Comparative Results

We compare the proposed method with five well-known methods including Viterbi algorithm (Viterbi) proposed in [4], Stroke Bounding Boxes method (SBB) proposed in [5], Connected Components Algorithm (CCA) proposed in [7], Project Profile Histogram (PPH) proposed in [6] and k-means proposed in [9]. Each of the compared methods is well-adjusted/trained to generate the best results.Both the correct rate and average segmentation time per character are reported and compared.

On each database, we select only text lines (strings) consisting of nonlinearly separable characters that are hard to segmented by vertical straight lines. On the IAM database, we obtain 156 strings comprising 964 characters. On the KAIST Hanja1 database, we obtain 233 strings consisting of 855 characters. On the HCL2000 database, we obtain 301 strings consisting of 1128 characters. On the SYSU database, we obtain 245 signatures consisting of 950 characters. The comparative results on four databases are listed in Table 1. As a overall comparison, Fig. 5 plots the comparative results on all strings collected from four databases according to the character types (English and Chinese) respectively.

Table 1. Comparing the correct rate (CR) and average segmentation time per character in seconds on four databases.

Algorithms	IAM		KAIST Hanja1		HCL2000		SYSU	
	CR(%)	Time(Sec.)	CR(%)	Time(Sec.)	CR(%)	Time(Sec.)	CR(%)	Time(Sec.)
Viterbi [4]	79.2	1.351	77.4	1.353	79.2	1.353	77.8	1.458
SBB [5]	82.7	1.271	81.3	1.221	81.3	1.241	82.9	1.624
CCA [7]	75.4	1.244	74.3	1.644	77.1	1.544	75.3	1.332
PPH [8]	73.5	1.115	72.6	1.334	70.8	1.417	72.6	1.749
k-means[9]	61.2	1.746	46.2	1.819	57.2	1.525	51.2	1.847
Our approach	85.4	1.102	89.2	1.162	88.9	1.189	89.7	1.025

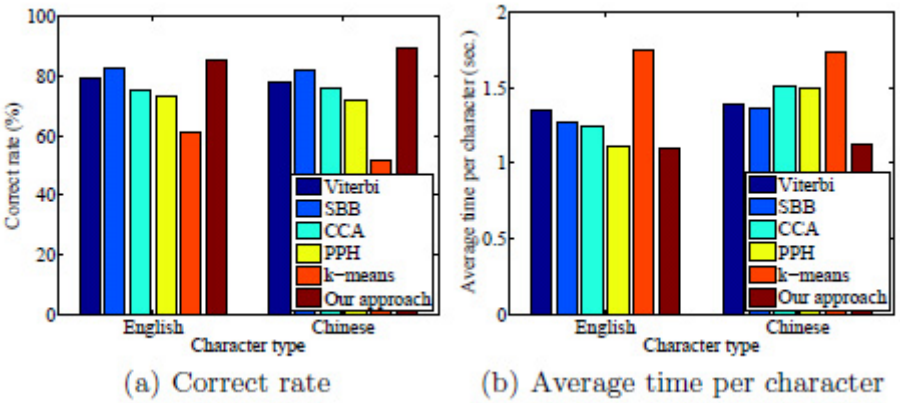


Fig. 5. Overall comparison in segmenting both English and Chinese characters.

By comparison, our approach is effective in segmenting both English and Chinese characters. The experimental results have demonstrated the effectiveness and efficiency of our approach.

5. Conclusions

In this paper, we present a new handwritten signature segmentation method, which can effectively segment nonlinearly separable characters. In the proposed approach, we first segment the entire text line into strokes, Then, the nonlinear Normalized cut (Ncut) clustering method is performed on this similarity matrix to obtain cluster labels for these strokes, through which the strokes are combined. Comparative results with some well-known methods on four testing databases demonstrate the effectiveness and efficiency of our approach.

Acknowledgments

This work was supported by the Technology Program of Yuexiu district of Guangzhou, and the Technology Program of Guangdong (2009B060700091).

References

[1] M. Bulacu and L. Schomaker. Text-independent writer identification and verification using textural and allographic features. IEEE Trans. Pattern Anal. Mach.Intell., 2007,29(4):701–717.
[2] G. Zhu, Y. Zheng, D. Doermann, and S. Jaeger. Signature detection and matching for document image retrieval. IEEE Trans. Pattern Anal. Mach. Intell., 2009,31(11):2015–2031,.
[3] M. Panagopoulos, C. Papaodysseus, P. Rousopoulos, D. Dafi, and S. Tracy. Automatic writer identification of ancient Greek inscriptions. IEEE Trans. Pattern Anal. Mach. Intell., 2009,31(8):1404–1414.
[4] Yi-Hong Tseng and Hsi-Jian Lee. Recognition-based handwritten Chinese character segmentation using a probabilistic Viterbi algorithm. Pattern Recognition Letters, 1999,20:791–806.
[5] Lin Yu Tseng and Rung Ching Chen. Segmenting handwritten Chinese characters based on heuristic merging of stroke bounding boxes and dynamic programming. Pattern Recognition Letters, 1998,19:963–973.
[6] T. Yamaguchi, T. Yoshikawa, T. Shinogi, S. Tsuruoka, and M. Teramoto. A segmentation method for touching Japanese handwritten characters based on connecting condition of lines. In Proc. of the 6th Int. Conf. Document Analysis and Recognition, 2001, 837–841.
[7] Lin-Wan Wang, Yang Yang, Bin Xie, and Yi Yang. Segmentation of handwritten Chinese characters based on connected components algorithm and penetration times algorithm. Information Technology, 2004,28(4):31–35.
[8] J. Shi and J. Malik. Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell., 2000,22(8):888–905.
[9] An-Bang Wang and Kuo-Chin Fan. Optical recognition of handwritten Chinese characters by hierarchical radical matching method. Pattern Recognition, 2001,34:15–35.
[10] J. MacQueen. Some methods for classification and analysis of multivariate observations. In Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, 1967, 1:281–297.
[11] Chang-Dong Wang, Jian-Huang Lai, and Jun-Yong Zhu. A conscience on-line learning approach for kernel-based clustering. In Proc. of the 10th Int. Conf. on Data Mining,2010,531–540.
[12] T. Y. Zhang and C. Y. Suen. A fast parallel algorithm forthinning digital patterns.Communications of the ACM, 1984,27(3):236–239.
[13] U.-V. Marti and H. Bunke. The IAM-database: an English sentence database for offline handwriting recognition. Int. Journal of Document Analysis and Recognition, 2002,5:39–46.
[14] In-Jung Kim and Jin-Hyung Kim. Statistical character structure modeling and its application to handwritten Chinese character recognition. IEEE Trans. Pattern Anal. Mach. Intell., 2003,25(11):1422–1436.

[15] Honggang Zhang, Jun Guo, Guang Chen, and Chunguang Li. HCL2000a large-scale handwritten Chinese character database for handwritten character recognition. In Proc. of the 10th Int. Conf. on Document Analysis and Recognition, 2009.