Journal of Zhejiang University SCIENCE
ISSN 1009-3095
http://www.zju.edu.cn/jzus
E-mail: jzus@zju.edu.cn

An approach to offline handwritten Chinese character recognition based on segment evaluation of adaptive duration*

LI Guo-hong (李国宏)†, SHI Peng-fei (施鹏飞)

(*Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai 200030, China*)

†E-mail: lgh0929@sjtu.edu.cn

Received May 7, 2003; revision accepted Aug. 18, 2003

Abstract: This paper presents a methodology for off-line handwritten Chinese character recognition based on mergence of consecutive segments of adaptive duration. The handwritten Chinese character string is partitioned into a sequence of consecutive segments, which are combined to implement dissimilarity evaluation within a sliding window whose durations are determined adaptively by the integration of shapes and context of evaluations. The average stroke width is estimated for the handwritten Chinese character string, and a set of candidate character segmentation boundaries is found by using the integration of pixel and stroke features. The final decisions on segmentation and recognition are made under minimal arithmetical mean dissimilarities. Experiments proved that the proposed approach of adaptive duration outperforms the method of fixed duration, and is very effective for the recognition of overlapped, broken, touched, loosely configured Chinese characters.

Key words: Handwritten Chinese character, Segmentation boundary, Segment, Duration
doi:10.1631/jzus.2004.1392 **Document code:** A **CLC number:** TP391.4

INTRODUCTION

Offline handwritten Chinese character reading by computer program is a complicated task. It is essential to separate a given character string correctly into the sequence of characters. Any failure or error in the segmentation step directly produces a negative effect on recognition (Lee and Lee, 1996).

The difficulty of handwritten Chinese character segmentation comes from the great variety of handwritings. In the case of hand-printed script (Fig.1a), segmentation is a relatively simple task. In the case of overlapped script (Fig.1b), broken character (Fig.1c), touched characters (Fig.1d), loosely configured character (Fig.1e), and mixed script (Fig.1f), segmentation is difficult. Overlapped, broken, touched and loosely configured characters are major causes of segmentation errors.

Previous approaches for Chinese character recognition mainly fall into straight segmentation and recognition method, wherein script is separated into a sequence of characters with each character discriminated respectively (Kahan *et al.*, 1987). Its limitation is its dependence on high accuracy of segmentation. But such high accuracy segmentation is very difficult for unconstrained handwritten Chinese characters; so that script segmentation and character recognition must be combined. Recent developments in handwritten English character recognition by recognition-based segmentation can be found in (Nafiz and Fatos, 2002; Favata, 2001). In this recognition-based method, a group of candi-

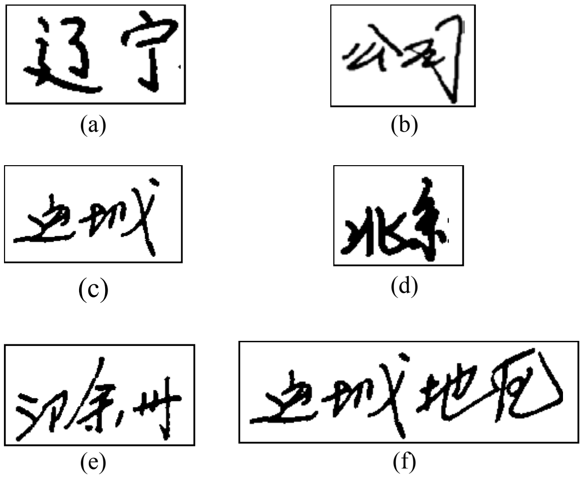
* Project (No. 60075007) supported by the National Natural Science Foundation of China

date segmentation boundaries are found, and the candidate boundaries are confirmed by recognition results. This method is more reasonable than the first one for handwritten character recognition, but it highly depends on the performance of the recognizer (Liang *et al.*, 1994).

This paper's proposed recognition-based methodology for handwritten Chinese character segmentation and recognition makes the best use of shape features and recognition results. For the mergence of consecutive segments, a method of variable duration for each character based on statistics for English characters is described in (Kim and Govindaraju, 1997). In this paper, the duration for each character is determined adaptively by shape and evaluation.

The set of candidate segmentation boundaries is crucial to handwritten Chinese character recognition. We make the following assumption: the correct segmentation boundaries are embedded in the set of candidate segmentation boundaries.

In order to verify the proposed method, the average stroke width is computed, then a set of candidate segmentation boundaries is explored, and segment-level evaluation based on combination of consecutive components of adaptive duration is described. A reduced directed graph is constructed according to the sequence of evaluation results. The final segmentation and recognition decisions are determined by searching the path of minimal arithmetical mean dissimilarity.



(a) (b) (c) (d) (e) (f)

Fig.1 Types of handwriting

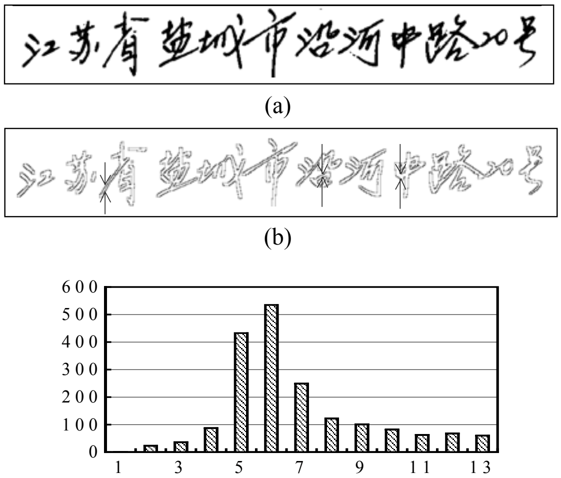
(a) Hand-printed; (b) Overlapped; (c) Broken; (d) Touched; (e) Loosely configured; (f) Mixed

ESTIMATION OF AVERAGE STROKE WIDTH

It is assumed that stroke width varies locally depending on writing device, paper texture and pressure within a script. Therefore, it is reasonable to get average stroke width and use it for subsequent image processing procedures in adaptive manner. We detect components of small size as noise, positions of small stroke width as candidate segmentation points, and shapes of some range of stroke width as components of characters.

To estimate the average stroke width of a given handwritten character string (as shown in Fig.2a), *MSW*, contours of that image are extracted. By tracing contours from the left-most column to the right-most column, the following distances (as shown in Fig.2b) for each column are computed: (1) distance between upper trace and lower trace of the outer contour, (2) distance between lower trace and upper trace of the inner contour, (3) distance between upper trace of inner contour and upper trace of the outer contour, and (4) distance between lower trace of inner contour and lower trace of the outer contour.

The histogram in Fig.2c shows the number of



(a) (b) (c)

Fig.2 Estimation of stroke width

(a) Original image of text line; (b) Contour from image of text line; (c) Histogram of stroke width occurrences of each distance value for the image of Fig.2b. The maximum of the histogram indicates an estimation of the average stroke width.

We define a set of candidate segmentation points

$$SPMSW=\{p|msw(p)<\alpha\cdot MSW\} \quad (1)$$

where $msw(p)$ is the stroke width at point p, and the coefficient, α ($0<\alpha<1$) is set empirically.

Generally, those stroke widths in points of a ligature are obviously smaller than average stroke width. Therefore, it is possible that some segmentation points between touched characters are embedded in *SPMSW*.

DETERMINATION OF CANDIDATE SEGMENTATION BOUNDARIES

To split handwritten Chinese characters into a sequence of consecutive segments, we apply a reduced nonlinear segmentation algorithm to the text image. The nonlinear segmentation algorithm (Lee and Lee, 1996; Nafiz and Fatos, 2002) performs the segmentation on the gray-scale image, but our reduced algorithm implements segmentation just on binary image.

First, the potential segmentation regions are explored by analyzing the variation locally on pixel and stroke projection profiles (the numbers of occurrences of black pixels and strokes in each column, respectively). Determination of the potential segmentation regions in a string image is implemented in the following steps:

(1) A column is labeled a non-segmentation-region if its stroke projection value is locally minimal, and its pixel projection value is locally maximal.

(2) For all unlabeled columns so far, those between two successive non-segmentation-regions belong to a segmentation region.

The potential segmentation regions carry candidate segmentation boundaries between characters. One candidate segmentation boundary at most in each potential segmentation region is found by the reduced nonlinear segmentation algorithm, which split the connected characters or components. Fig.3b shows the candidate segmentation boundaries from reduced nonlinear segmentation algorithm.

SEGMENT-LEVEL EVALUATION

Those candidate segmentation boundaries explored by the above reduced nonlinear segmentation algorithm often partition a single character into several components. In order to confirm the character segmentation boundaries and recognition results, the scheme based on evaluations of merged segments of adaptive durations is undertaken. The evaluations are represented as directed evaluation graph, and the optimal character segmentation boundaries are confirmed by exploring a minimal mean cost path in the evaluation graph.

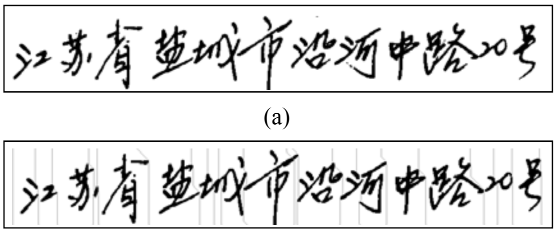
The component between two successive candidate segmentation boundaries is defined as a segment. The combinations of consecutive segments are fed to recognizer, and a series of dissimilarity evaluation results is obtained for a sequence of combinations.

Suppose the set of candidate segmentation boundaries can be denoted as

$$SB=\{sp_0, sp_1, \ldots, sp_N\} \quad (2)$$

where N is the number of segments in character string and sp_k is the kth candidate segmentation boundary.

Suppose a handwritten Chinese character string



(a)

(b)

Fig.3 Result of pre-segmentation

(a) Original text line image; (b) Candidate boundaries by nonlinear segmentation algorithm can be denoted as

$$CC=\{s_0,s_1,\ldots,s_{N-1}\} \tag{3}$$

where s_k is the kth segment, and s_k denotes the component between candidate boundaries sp_k and sp_{k+1}.

Suppose the original text line can be referred to as

$$TL=\{ c_0,c_1,\ldots,c_{M-1}\} \tag{4}$$

where M is the number of characters in character string and c_k is the kth character.

Our endeavor is to find such a subset from the set SB whose elements separate the text image into M segments that the arithmetical mean of evaluations of those segments is minimal.

We define the notion of merged segments as

$$cs(s,e)=\{s_s\oplus s_{s+1}\oplus\ldots\oplus s_{e-1}\} \tag{5}$$

where s is the start segmentation boundary of merged segments, and e is the end boundary of merged segments.

Generally, an image of merged consecutive segments of width csw and height csh is potentially a character pattern when $csw<T_{h1}\cdot csh$. The coefficient T_{h1} is empirically set, generally the maximal from statistics. Therefore, the mergence of segments from s is terminated when $csw\geq T_{h1}\cdot csh$. An image of segment of mass csm could be thought of as a block of noise when $csm<T_m\cdot MSW^2$. Therefore, an image of segment of mass csm can be thought of as a character pattern or one component of character pattern when $csm\geq T_m\cdot MSW^2$. An image of segment of width csw and height csh is potentially a character pattern when $csw>T_{h2}\cdot MSW$ or $csw>T_{h3}\cdot MSW$. Therefore, image of merged consecutive segments of width csw and height csh would not be considered as a character pattern when $csw\leq T_{h2}\cdot MSW$ and $csh\leq T_{h3}\cdot MSW$. The coefficients T_{h2} and T_{h3} are empirically set, generally the minimal from statistics.

We define the evaluation result of merged segments as

$$\boldsymbol{D}_s^e = D(cs(s,e)) \tag{6}$$

where $\boldsymbol{D}_s^e = \begin{bmatrix} d_s^e \\ r_s^e \end{bmatrix}$, and r_s^e is the resulting character corresponding to the minimal dissimilarity, d_s^e.

Generally, an image of merged consecutive segments is potentially a character pattern when $d_s^e < T_d$. The coefficient T_d is empirically set, generally the maximal from statistics. The merge of segments from s to e is not considered a candidate character pattern when $d_s^e \geq T_d$. An image of merged consecutive segments $cs(s,e)$ would not be thought of as character pattern when $d_s^e > T_{d1}d_s^{e-1}$ or $d_s^{e+1} < T_{d2}d_s^e$. The coefficients T_{d1} (>1) and T_{d2} (<1) are empirically set.

Then we get the set of recognition results of the sequence of merged segments

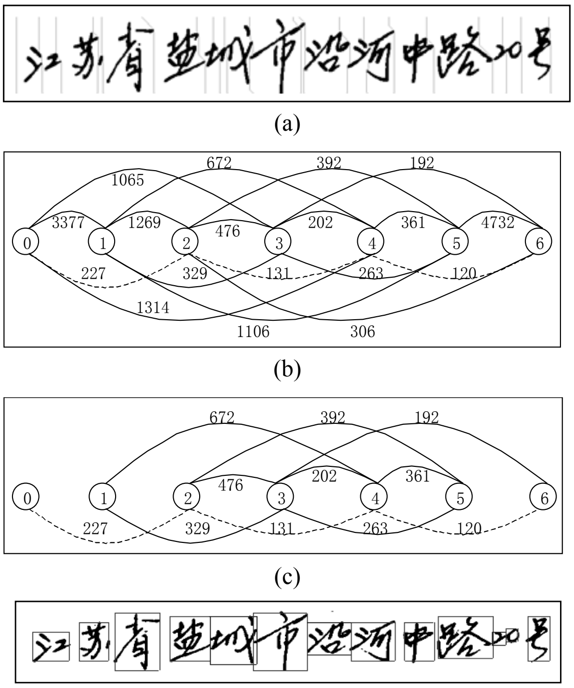
$$\begin{aligned}\boldsymbol{SD} = \{&\boldsymbol{D}_0^1, \boldsymbol{D}_0^2, \cdots \boldsymbol{D}_0^{SW_0}, \boldsymbol{D}_1^2, \boldsymbol{D}_1^3, \cdots \\ &\boldsymbol{D}_1^{SW_1+1}, \cdots, \boldsymbol{D}_{N-2}^{N-1}, \boldsymbol{D}_{N-2}^{N}, \boldsymbol{D}_{N-1}^{N}\}\end{aligned} \tag{7}$$

Where SW_i denotes the maximum of consecutive segments from starting ith segment to be combined.

Fig.4c represents the evaluation results of merged segments of adaptive duration in the form of directed graph. The nodes denote candidate segmentation boundaries, the edges signify the merged segments, and the weights indicate evaluation results, dissimilarities. Fig.4b shows the evaluation results using fixed duration. Obviously, the graph of adaptive duration accelerates the searching process for final segmentation and recognition decision.

CHARACTER SEGMENTATION AND RECOGNITION

The final decisions on segmentation and recognition are made under minimal arithmetical mean dissimilarity that is more reasonable than minimal accumulated dissimilarity. We define a path as such a set of edges that the ending node can be reached



(a)

(b)

(c)

(d)

Fig.4 Decision on segmentation and recognition
(a) Candidate segmentation boundaries; (b) Evaluation graph of fixed duration; (c) Evaluation graph of adaptive duration; (d) Final segmentation decision

from starting node by consecutively traversing them. The sequence of possible paths on the evaluation graph can be denoted as

$$P=\{p_k|k=0,1,\ldots,PN-1\} \tag{8}$$

Where p_k is a path and PN is the number of possible paths in the evaluation graph. Let d_i^j be the dissimilarity from node i to node j, $E(p_k)$ be a set of edges on the path p_k, and $|E(p_k)|$ be the number of edges in $E(p_k)$. The mean dissimilarity of path p_k can be computed as follows:

$$d(p_k)=\frac{\sum\limits_{<i,j>\in E(p_k)} d_i^j}{|E(p_k)|} \tag{9}$$

The optimal character segmentation path can be confirmed by searching the path of minimal mean dissimilarity,

$$p_m=\arg\min_{p_k}\{d(p_k)\} \tag{10}$$

The dashed edges shown in Fig.4b and Fig.4c indicate the optimal character segmentation paths. Based on the minimal mean dissimilarity segmentation path, the segmentation boundaries (i.e. the nodes on the optimal segmentation path) are picked up. The bounding boxes of characters between two successive segmentation boundaries are drawn with straight lines (Fig.4d). The evaluation results on merged segments in bounding boxes turn out the final recognition results.

Suppose the minimal mean dissimilarity segmentation path can be referred to as:

$$p_m=\{n_0,n_1,\ldots,n_M\} \tag{11}$$

where n_i is the ith node on the path of minimal mean dissimilarity.

The values of dissimilarity $d_{n_i}^{n_{i+1}}$ and character $r_{n_i}^{n_{i+1}}$ are computed as follows:

$$d_{n_i}^{n_{i+1}}=\min_{0\le k\le VN} d(V(cs(n_i,n_{i+1}),VT_k) \tag{12}$$

$$r_{n_i}^{n_{i+1}}=\arg\min_{0\le k\le VN} d(V(cs(n_i,n_{i+1})),VT_k) \tag{13}$$

where $d(\cdot)$ indicates dissimilarity computation between two feature vector, operator $V(\cdot)$ performs feature extraction for merged segments, VT_k is the feature vector of the kth character template, and VN is the number of character templates.

Fig.4d shows the final decision on segmentation.

EXPERIMENTAL RESULTS AND DISCUSSION

Experiments for handwritten Chinese character string segmentation and recognition were carried out with text line images from 500 Chinese mail address lines. All the images are binary. Those lines contain many overlapped, touched, broken, and loosely configured characters. The input images to these experiments are binary without any operations of smoothing and correction.

Given a script line image, the result of the candidate segmentation boundaries is considered to be correct if there exists a path from source to sink in the word graph. There are almost no cases not completely partitioned in those text line images. The final character segmentation and recognition decisions are highly inter-related, and the correct recognition accuracy of the top 5 candidate characters for this kind of script lines is 93%. The correct recognition accuracy of address words is greater than 90%.

Because of the variety of character shapes from handwriting styles, we make decision about segmentation and recognition based on the evaluation of merged segments, instead of the characterization of such character shapes as aspect ratio and width. The proposed strategy made a recognition solution for not only touched and overlapped characters (as shown in Fig.5 and Fig.4), but also the combination of broken and loosely configured characters (as shown in Fig.4). The methodology is effective for segmentation and recognition decision on the above characters, and greatly improved the recognition performance.

Obviously, the proposed strategy of adaptive duration requires much less expensive computation resource than that of fixed duration, and improves the performance of segmentation and recognition because of the optimization rule of minimal mean dissimilarity.

There are still main sources of segmentation failure to be identified: (1) multi-touched character string, (2) the accuracy of evaluations. The variation of handwritings decreases the accuracy of evaluation, which negatively affects the final recognition decision.

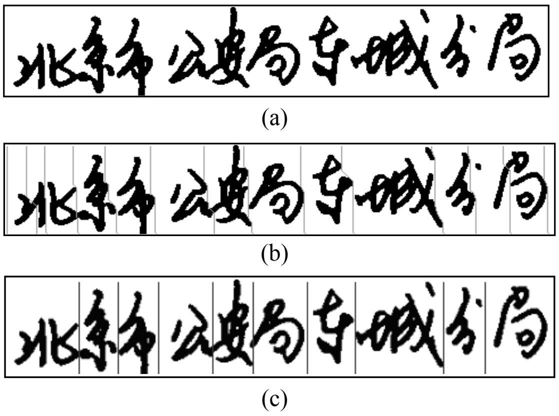


Fig.5 Experimental results

(a) Text line image; (b) Candidate segmentation boundaries; (c) Final segmentation and recognition result

CONCLUSION

We encountered difficulties from broken, overlapped, touched, and loosely configured characters for handwritten Chinese character segmentation and recognition. In this paper, we proposed a new strategy for handwritten Chinese character string segmentation and recognition that combines consecutive segments within a sliding window of adaptive duration. The segmentation and recognition are highly inter-related. We have proved that many of errors from these problems can be tackled by the proposed approach.

There are still some ways to improve the segmentation and recognition: (1) exploiting context information from a lexicon, and (2) segmenting the string into a sequence of finer segments will certainly improve the segmentation, while bringing about more expenses in computation.

References

Favata, J.T., 2001. Offline general handwritten word recognition using an approximate BEAM matching algorithm. *IEEE Trans. on Pattern Recognition and Machine Intelligence*, **23**(9):1009-1021.

Kahan, S., Pavlidis, T., Baird, H.S., 1987. On the recognition of printed characters of any fonts and sizes. *IEEE Trans. on Pattern Recognition and Machine Intelligence*, **9**(2):274-288.

Kim, G., Govindaraju, V., 1997. A lexicon driven approach to handwritten word recognition for real-time applications. *IEEE Trans. on Pattern Recognition and Machine Intelligence*, **19**(4):366-379.

Lee, S.W., Lee, D.J., 1996. A new methodology for gray-scale character segmentation and recognition. *IEEE Trans. on Pattern Recognition and Machine Intelligence*, **18**(10):1045-1050.

Liang, S., Shridhar, M., Ahmadi, M., 1994. Segmentation of touching characters in printed document recognition. *Pattern Recognition*, **27**(6):825-840.

Nafiz, A., Fatos, T.Y., 2002. Optical character recognition for cursive handwriting. *IEEE Trans. on Pattern Recognition and Machine Intelligence*, **24**(6):801-813.