HAL
archives-ouvertes.fr

Recognition-based segmentation of Nom characters from body text regions of stele images using area Voronoi diagram

Thai V. Hoang, Salvatore Tabbone, Ngoc-Yen Pham

▶ **To cite this version:**

Thai V. Hoang, Salvatore Tabbone, Ngoc-Yen Pham. Recognition-based segmentation of Nom characters from body text regions of stele images using area Voronoi diagram. Xiaoyi Jiang and Nikolai Petkov. International Conference on Computer Analysis of Images and Patterns - CAIP'2009, Sep 2009, Munster, Germany. Springer-Verlag, 5702, pp.205-212, 2009, Lecture Notes in Computer Science. <http://www.springerlink.com/content/k375k70016407p6v/?p=ffe4abba883b40fb91039d0c110af295

pi=9>. <10.1007/978-3-642-03767-2>. <inria-00430806>

HAL Id: inria-00430806
https://hal.inria.fr/inria-00430806

Submitted on 10 Nov 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Recognition-based Segmentation of Nom Characters from Body Text Regions of Stele Images Using Area Voronoi Diagram

Thai V. Hoang*[1,2], Salvatore Tabbone[2], and Ngoc-Yen Pham[1]

[1] MICA Research Center, Hanoi University of Technology, Hanoi, Vietnam
[2] Université Nancy 2, LORIA, UMR 7503, 54506 Vandoeuvre-lès-Nancy, France
{vanthai.hoang,tabbone}@loria.fr,ngoc-yen.pham@mica.edu.vn

Abstract. Segmentation of Nom characters from body text regions of stele images is a challenging problem due to the confusing spatial distribution of the connected components composing these characters. In this paper, for each vertical text line, area Voronoi diagram is employed to represent the neighborhood of the connected components and Voronoi edges are used as nonlinear segmentation hypotheses. Characters are then segmented by selecting appropriate adjacent Voronoi regions. For this purpose, we utilize the information about the horizontal overlap of connected components and the recognition distances of candidate characters provided by an OCR engine. Experimental results show that the proposed method is highly accurate and robust to various types of stele.

Key words: Recognition-based segmentation, area Voronoi diagram, stele images, Nom characters, horizontal overlap, segmentation graph

1 Introduction

Stone steles in Vietnam usually contain Nom characters, a derivative of Chinese which was used before the 20th century describing important historical events. Today, the exploitation of these steles is necessary to better understand the history and form a solid base for future development. Automatic processing of stele images composes of three sub-problems: extraction of body text regions, segmentation of Nom characters, and finally representation of these Nom characters in a database allowing search for information. The first one has been tackled using the information about the thickness of connected components [1]. Fig. 1(a)-1(b) show an example of a stele image and its extracted body text region respectively.

As shown in Fig. 1(b), Nom characters are engraved on stone steles in vertical text line from right to left and each character may be composed of several connected components. In each text line, the gaps between characters is indistinguishable from the gaps between connected components belonging to one

* This work is supported by CNRS and EFEO under the framework of an International Scientific Cooperation Program (PICS: 2007-2009).

(a) (b)

Fig. 1. A stele image (a) and its extracted body text region (b)

character. This makes segmentation of Nom characters a challenging problem to be dealt with in this paper.

There exists many methods in literature for character segmentation. Lu [2], Lu and Shridhar [3] reviewed methods for character segmentation in machine printed documents and handwritten works. Casey and Lecolinet [4] classified existing methods into three strategies: *dissection*, *recognition-based*, and *holistic*. The first strategy decomposes the image into a sequence of sub-images using general features like character height, width and white space between characters. In the second strategy, the system searches the image for components that match classes in its alphabet. The third strategy seeks to recognize the words as a whole avoiding the need to segment into characters. This strategy is inappropriate for Nom characters as each Nom character has its own meaning.

Tseng and Chen [5] proposed a method for Chinese character segmentation by first generating bounding boxes for character strokes then using knowledge-based merging operations to merge these bounding boxes into candidate boxes and finally applying dynamic programming algorithm to determine optimal segmentation paths. However, the assumption of similarity on character sizes makes this method unsuitable for Nom characters written vertically. Viterbi algorithm and background thinning method were used in [6, 7] to locate nonlinear segmentation hypotheses separating handwritten Chinese characters. These methods are also inappropriate for Nom characters as there may exist horizontal gaps inside a character separating its connected components and connected components from neighboring characters on the same line may horizontally overlap.

Area Voronoi diagram has been used by some researchers for document image analysis. For example, Kise et al. [8] and Lu et al. [9] used area Voronoi diagram for page segmentation and word grouping in document images respectively using the distance and area ratio between neighboring connected components.

However, these methods work only for alphanumeric documents in which each character is represented as a connected components.

In this paper, we propose an efficient method combining *dissection* and *recognition-based* strategies. For each vertical text line extracted from the body text region using vertical projection profile, area Voronoi diagram is employed to represent the neighborhood of connected components and Voronoi edges are used as nonlinear segmentation hypotheses. Adjacent Voronoi regions are first grouped using the information about the horizontal overlap of connected components. The remaining Voronoi edges are used as vertices in a segmentation graph in which the arcs' weights are the recognition distances of the corresponding candidate characters using an OCR engine. The vertices in the shortest path detected from the segmentation graph represent the optimal segmentation paths.

The remainder of this paper is organized as follows. Section 2 briefly gives a basic definition of area Voronoi diagram. Section 3 presents a method to group adjacent Voronoi regions using horizontal overlap of connected components. Section 4 describes the details of the algorithm determining optimal segmentation paths using recognition distances of candidate characters. Experimental results are given in Section 5, and finally conclusions are drawn in Section 6.

2 Area Voronoi Diagram

Let $G = \{g_1, \ldots, g_n\}$ be a set of non-overlapping connected components in the two dimensional plane, and let $d(p, g_i)$ be the Euclidean distance between a point p and a connected component g_i defined by $d(p, g_i) = \min_{q \in g_i} d(p, q)$, then Voronoi region $V(g_i)$ and area Voronoi diagram $V(G)$ are defined by:

$$V(g_i) = \{p \mid d(p, g_i) \leq d(p, g_j), \forall j \neq i\}$$
$$V(G) = \{V(g_1), \ldots, V(g_n)\}$$

The Voronoi region of each image component corresponds to a portion of the two dimensional plane. It consists of the points from which the distance to the corresponding component is less than or equal to the distance to any other image components. The boundaries of Voronoi regions, which are always curves, are called *Voronoi edges*.

To construct area Voronoi diagram, we utilize the approach represented in [10] that first labels the image components and then applies morphological operations to expand their boundaries until two expanding labels are met. Fig. 2(a)-2(d) show the area Voronoi diagram for a column text line extracted from the body text region in Fig. 1(b). Original text line I is given in Fig. 2(a). Fig. 2(b) demonstrates the Euclidean distance map in which the gray value of each pixel is proportional to the distance from that pixel to the nearest connected component. The area Voronoi diagram V of I is shown in Fig. 2(c). Voronoi regions with their corresponding connected components are given in Fig. 2(d).

As shown in Fig. 2(d), each connected component is represented by one Voronoi region and Voronoi edges can be used as nonlinear segmentation hypotheses. The process of character segmentation is then considered as the process

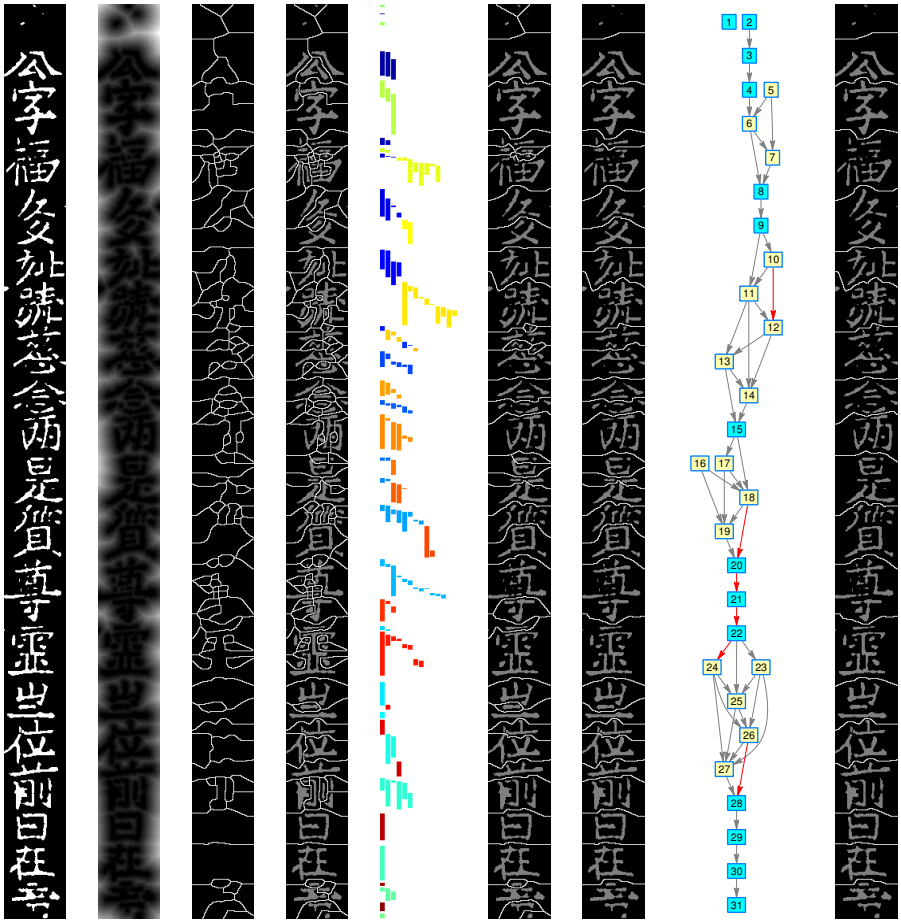


Fig. 2. Steps in segmenting Nom characters from one vertical text line

of grouping adjacent Voronoi regions representing one character. In this paper, we propose to group adjacent Voronoi regions in two steps:

- *Step 1*: Voronoi regions are first grouped using the criteria based on the degree of horizontal overlap of their corresponding connected components. We argue that connected components from one vertical text line overlapped horizontally to a certain degree should belong to one character. This algorithm is described in Section 3.
- *Step 2*: As each Nom character may be composed of several connected components and there may exist horizontal gaps between these connected components, the above algorithm does not guarantee a grouped Voronoi region for each character. We further group Voronoi regions using the recognition distances of candidate characters in Section 4.

3 Grouping of Voronoi Regions Using Horizontal Overlapping Profile of Connected Components

In order to calculate the degree of horizontal overlap of two connected components, we define T_i and B_i as the top and bottom coordinates of the bounding box of the connected component g_i ($B_i > T_i$). The degree of horizontal overlap VO_{ij} of two connected components g_i and g_j is calculated as:

$$VO_{ij} = \frac{max\{min(B_i - T_j, B_j - T_i), 0\}}{min(B_i - T_i, B_j - T_j)} \tag{1}$$

The numerator of (1) is interpreted as the length of the overlapping segment and VO_{ij} is the proportion of the shorter connected component being overlapped. Thus two connected components g_i and g_j which have $VO_{ij} \geq VO_{thr}$ are considered as belonging to one character and their corresponding Voronoi regions are grouped. Fig. 2(e) provides the horizontal projections of the bounding boxes of connected components in Fig. 2(a) with each vertical line corresponds to one connected component. The grouped Voronoi regions are shown in Fig. 2(f) and adjacent lines correspond to each grouped Voronoi region are labeled using the same color in Fig. 2(e).

By observation of Fig. 2(d) we realize that if each Nom character is represented by one group of Voronoi regions, these grouped region should span from the left border of the text line to its right border. From this viewpoint, those Voronoi regions that do not span from left to right in Fig. 2(f) need to be further grouped to one of its adjacent regions. We propose to use the distance d_{ij} between neighboring connected components g_i and g_j as the criterion of grouping:

$$d_{ij} = \min_{p_i \in g_i, p_j \in g_j} d(p_i, p_j)$$

where $d(p_i, p_j)$ is the Euclidean distance between p_i and p_j. Thus, for each Voronoi region i to be further grouped, we select a region j from a set of its adjacent regions D by $j = \text{argmin}_{k \in D}\, d_{ik}$ and then group region i with region j. The resulting grouped Voronoi regions are provided in Fig. 2(g).

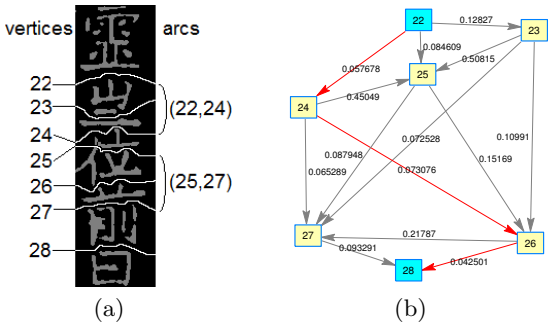
4 Recognition-based Determination of Optimal Segmentation Paths

The validity of segmentation hypotheses in Fig. 2(g) is verified by feeding candidate characters into an OCR engine and using their recognition distances to determine the optimal segmentation paths. A segmentation graph is constructed by using the segmentation hypotheses as its vertices and recognition distance of the candidate character corresponding to vertex i and vertex j as the weight of the arc connecting i and j. The segmentation graph in Fig. 2(h) has 31 vertices corresponding to 29 segmentation hypotheses in Fig. 2(g) plus the top and bottom lines. Assuming that a lower value of recognition distance corresponds to higher confidence of the OCR engine in the candidate character, the optimal

segmentation paths are thus determined by finding the shortest path in the segmentation graph. For the graph in Fig. 2(h), we need to find the shortest path from vertex 1 to 31.

In updating the weights of the graph, instead of feeding all candidate characters into the OCR engine, we only feed candidate characters that have the character-like feature $H \leq H_{thr}W$ where H and W are the height and width of the candidate character. By doing this, candidate character covering at least two Nom characters are mostly discarded. The arcs shown in Fig. 2(h) correspond to all candidate characters that have the character-like feature.

As there exists no OCR engine for Nom characters, we employ a Chinese OCR engine [11] admitting that not all Nom characters are recognizable by the engine. However, as the proportion of Nom characters that are not Chinese characters are small, the shortest path algorithm can overcome the case in which one non-Chinese character lies between two Chinese characters. An example of a line segment of Nom characters along with its segmentation graph are given in Fig. 3(b) and 3(a) respectively. The optimal segmentation paths in this case contains vertices $\{22, 24, 26, 28\}$.



(a) (b)

Fig. 3. (a) A line segment of Fig. 2(g), (b) Its corresponding segmentation graph

For the full segmentation graph in Fig. 2(h), directly applying a shortest path algorithm to the may result in error due to the inappropriateness of the OCR engine, we propose here a three-steps algorithm to find the shortest path of the segmentation graph:

- Find cut vertices (sky blue vertices in Fig. 2(h)): a cut vertex is a vertex the removal of which disconnect the remaining graph, the shortest path of the graph must contain cut vertices.
- Find arcs having corresponding candidate characters of high confidence (read arcs in Fig. 2(h)): candidate characters that have recognition distances less than RD_{thr} are said to be of high confidence and the arcs corresponding to these characters should be included in the shortest path.
- Find the shortest paths of the remaining subgraphs.

The final optimal segmentation paths are shown in Fig. 2(i).

5 Experimental Results

In order to determine the appropriate values of VO_{thr} and H_{thr}, a learning set of 16 Nom text lines containing 280 characters have been used for the evaluation of the proposed algorithm's accuracy at different threshold values. According to Fig. 4, the values of VO_{thr} and H_{thr} are selected as 0.4 and 1.25 respectively corresponding to the maxima of the curves. The value of RD_{thr}, which depends on the OCR engine, is selected experimentally as 0.06.

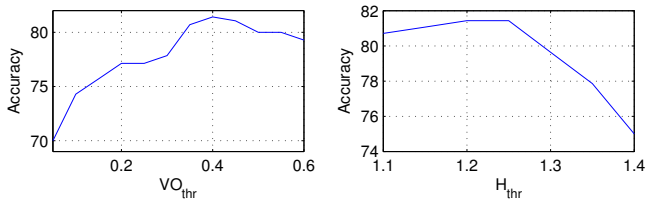


Fig. 4. Algorithm's accuracy at different values of VO_{thr} and H_{thr}

Out of 205 stele images provided by EFEO (The French School of Asian Studies) for evaluation, only 40 images are eligible for the proposed algorithm. The remaining images are regarded as insufficient because of their too noisy body text regions or their poor resolution. We have randomly selected 20 stele images containing 4998 Nom characters for experiment. The ground truth for these characters are defined by hand. Table 1 summarizes the experimental results. The accuracy which is defined as the percentage of characters that are segmented correctly has the value 89.14%. There are two sources of error: one is missing and the other is incorrect grouping. The error of missing concerns with characters that are classified as background noise. In incorrect grouping, each segmented character is not composed of all the connected components from one Nom character, its connected components may come from background noise or neighboring Nom characters.

Table 1. Performance of Nom character segmentation

	EFEO database
Accuracy(%)	89.14
Missing(%)	0.84
Incorrect grouping(%)	10.02

A considerable amount of error comes from the steles which contain characters of various sizes and curved text lines in their body text regions. The layout of these steles thus cannot be aligned into vertical straight lines. This results

in error in the extraction of text lines from body text regions using vertical projection profile and consequently the segmented characters will be inaccurate.

6 Conclusions

In this paper, area Voronoi diagram has demonstrated to be effective in representing the neighborhood of connected components in digital images. Voronoi edges then function as nonlinear segmentation hypotheses that need to be validated. Adjacent Voronoi regions have been first grouped using the information about the vertical overlap of connected components. The remaining Voronoi edges are used as vertices in a segmentation graph in which the weight of each arc is the recognition distance of the corresponding candidate character provided by an OCR engine. The vertices in the shortest path of the segmentation graph represent the optimal segmentation paths. Experimental results on a number of stele images show that the proposed method is highly accurate. Further work will employ a curved text line extraction algorithm [12] to increase the accuracy and represent each Nom character in a database for later retrieval. Moreover, poor-resolution images will be re-scanned and noise in body text regions will be removed to make this method more applicable.

References

1. Hoang, T.V., Tabbone, S., Pham, N.Y.: Extraction of Nom text regions from stele images using area Voronoi diagram. In: Proceedings of ICDAR 2009. (*to appear*)
2. Lu, Y.: Machine printed character segmentation - An overview. Pattern Recognition **28**(1) (1995) 67–80
3. Lu, Y., Shridhar, M.: Character segmentation in handwritten words - An overview. Pattern Recognition **29**(1) (1996) 77–96
4. Casey, R.G., Lecolinet, E.: A survey of methods and strategies in character segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **18**(7) (1996) 690–706
5. Tseng, L.Y., Chen, R.C.: Segmenting handwritten Chinese characters based on heuristic merging of stroke bounding boxes and dynamic programming. Pattern Recognition Letters **19**(10) (1998) 963–973
6. Tseng, Y.H., Lee, H.J.: Recognition-based handwritten Chinese character segmentation using a probabilistic Viterbi algorithm. Pattern Recognition Letters **20**(8) (1999) 791–806
7. Zhao, S., Chi, Z., Shi, P., Yan, H.: Two-stage segmentation of unconstrained handwritten Chinese characters. Pattern Recognition **36**(1) (2003) 145–156
8. Kise, K., Sato, A., Iwata, M.: Segmentation of page images using the area Voronoi diagram. Comput. Vis. Image Underst. **70**(3) (1998) 370–382
9. Lu, Y., Wang, Z., Tan, C.L.: Word grouping in document images based on Voronoi tessellation. In Marinai, S., Dengel, A., eds.: Document Analysis Systems. Volume 3163 of Lecture Notes in Computer Science., Springer (2004) 147–157
10. Lu, Y., Xiao, C., Tan, C.L.: Constructing area Voronoi diagram based on direct calculation of the Freeman code of expanded contours. IJPRAI **21**(5) (2007) 947–960

11. Léonard, J.: COCR2: A small experimental Chinese OCR. *Available at http://users.belgacom.net/chardic/cocr2.html*
12. Likforman-Sulem, L., Zahour, A., Taconet, B.: Text line segmentation of historical documents: a survey. IJDAR **9**(2-4) (2007) 123–138