Development of Nom Character Segmentation for Collecting Patterns from Historical Document Pages

Truyen Van Phan
Tokyo Univ. of Agri. & Tech.
Naka-cho 2-24-16, Koganei
Tokyo, Japan
truyenvanphan@gmail.com

Bilan Zhu
Tokyo Univ. of Agri. & Tech.
Naka-cho 2-24-16, Koganei
Tokyo, Japan
zhubilan@cc.tuat.ac.jp

Masaki Nakagawa
Tokyo Univ. of Agri. & Tech.
Naka-cho 2-24-16, Koganei
Tokyo, Japan
nakagawa@cc.tuat.ac.jp

ABSTRACT
In this paper, we present the first effort in preprocessing and character segmentation on digitized Nom document pages toward their digital archiving. Nom is an ideographic script to represent Vietnamese, used from the 10th century to 20th century. Because of various complex layouts, we propose an efficient method based on connected component analysis for extraction of characters from images. The area Voronoi diagram is then employed to represent the neighborhood and boundary of connected components. Based on this representation, each character can be considered as a group of extracted adjacent Voronoi regions. To improve the performance of segmentation, we use the recursive x-y cut method to segment separated regions. We evaluate the performance of this method on several pages in different layouts. The results confirm that the method is effective for character segmentation in Nom documents.

Categories and Subject Descriptors
I.4.6 [**Image Processing and Computer Vision**]: Segmentation – *edge and feature detection, region growing, partitioning.*

General Terms
Experimentation

Keywords
Segmentation, Chu Nom, historical document, Area Voronoi Diagram, connected component analysis, document image analysis, Vietnamese historical document.

1. INTRODUCTION
Around the early part of the 10th century, Vietnamese scholars invented Nom as an ideographic script to represent the Vietnamese language. From the 10th century and into the 20th, much of Vietnamese literature, philosophy, history, law, and so on were written in the Nom script. However, Nom is now entirely obsolete. Today, less than 100 scholars world-wide can read Nom. Much of history of Vietnam is inaccessible to the 80 million speakers of the language. Therefore, it is extremely necessary to preserve and utilize this cultural heritage. Vietnamese agencies and world-wide foundations have been collected thousands volumes of Han Nom textbooks. Recently, owing to the information technology in preserving, managing, researching, and discovering Han Nom resources, over 4,000 books have been scanned to digital images as shown in Figure 1 [1].

In order to utilize this huge resource of knowledge, however, they must be indexed, annotated, and hopefully recognized and translated into current Vietnamese. Since the translation is most difficult and time-consuming, with a limited and even decreasing number of experts, the process cannot be done in a short period of time. Obviously, document recognition techniques can speed up the digitalization process.

The same problem has been recognized in Korea and a digitalizing scheme of handwritten Hanja historical documents has been proposed by Kim et al. [2]. This approach seems to be a good reference for digitalizing Han Nom. Unfortunately, there is no OCR for Nom. We have to start from collecting patterns. In order to collect them from Nom document images, we need to develop a reliable segmentation method which helps us collect sample patterns and also builds an OCR system.

Figure 1. Example of image for Nom historical document

There exist many methods in literature for character segmentation. Tseng and Chen [3] proposed a method for Chinese character segmentation by first generating bounding boxes for character strokes then using knowledge-based merging

HIP '11, September 16 - September 17 2011, Beijing, China
Copyright 2011 ACM 978-1-4503-0916-5/11/09 $10.00.

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 [4][5] to locate nonlinear segmentation hypotheses separating handwritten Chinese characters. These methods are inappropriate for Nom characters, however, since they are just methods which segment character from text-lines or even they are recognition-based ones.

The area Voronoi diagram has been used by some researchers for document image analysis. For example, Kise et al. [6] and Lu et al. [7] used the area Voronoi diagram for page segmentation and word grouping in document images respectively. However, these methods work only for alphanumeric documents in which each character is represented as a connected component.

However, in Nom historical documents, there are many kinds of layout types as known as non-Manhattan possibly, and various sizes of character in different documents. In this paper, we propose an effective method that is unconstrained by document layout and character size. The area Voronoi diagram is employed to represent the neighborhood of connected components. Adjacent Voronoi regions are then grouped by using information of nearest neighbor distance.

The remainder of this paper is organized as follows. Section II introduces overview of the proposed system. Section III presents the preprocessing process. Section IV briefly defines the area Voronoi diagram and how to construct it. Section V describes the details of the segmentation method. Experiment results are reported in Section VI. Finally some conclusions are drawn in Section VII.

2. SYSTEM OVERVIEW

The proposed system for segmenting Nom characters from document images consists of three main steps: 1) preprocessing, 2) area Voronoi diagram construction, and 3) grouping Voronoi regions. Overview of the system is drawn in Figure 2.

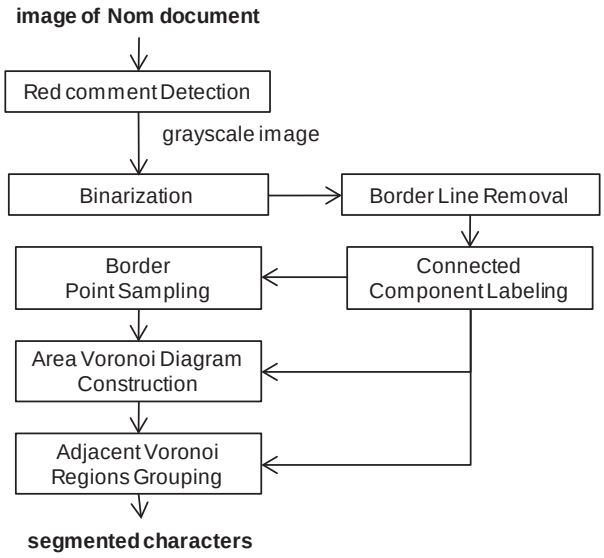


Figure 2. Overview of system

As shown in Figure 1, many characters in the Nom documents are accompanied with red comments. In preprocessing step, we turn their color to light gray lower than the binarization threshold. The image is then binarized by an effective method for badly degraded historical document [8] with comments and other noises removed. Finally, the border line covering the text region is removed.

In the next step, in order to construct the area Voronoi diagram, we must obtain points which are then used as Voronoi points to generate the diagram. We implement a fast algorithm for labeling connected components and sampling border points concurrently. In some documents, the characters may be written neatly in rows and columns, we can use x-y cut method to divide document into smaller regions firstly and then we construct diagram in each region.

After all, based on information of connected components, neighbor relations and distance between components, we group adjacent Voronoi regions to form character patterns.

3. PREPROCESSING

3.1 Red Comment Removal

We recognized that in most of documents, there are many red comments beside characters. If we binarized the document image immediately, there comments may make noise. For this reason, firstly we must remove or blur them.

In image processing, there are many methods which allow filtering pixels depending on their color values, such as color filtering, channel filtering, HSL filtering, Euclidean color filtering, etc. These image processing filters may be used to keep pixels, whose color falls inside or outside of a specified range, and fill the rest of pixels with a specified color. In this system, we use the method of color filtering in the HSL color space because of its effect to detect red color in our documents. We specified the range of hue, saturation, luminance as (300, 50), (0.3, 1.0), (0.3, 1) respectively to detect red color of comments. If pixel's color is inside of this specified range, it is turned to light gray which will be deleted by the following binarization.

The figure 3(b) illustrates red comment removed result from the document image of Figure 3(a).

3.2 Binarization

Since historical document collections are usually in poor quality caused by degradations include shadows, non-uniform illumination, ink seeping, etc. The situation is the same as in the Nom documents. Thus, we need an effective and efficient method to handle this problem.

In recent years, many document image binarization techniques have been reported. Specially, among them, a simple but effective method for historical documents has been proposed by Su et al. [8]. The technique makes use of the image contrast that is defined by the local image maximum and minimum. In particular, the processing of document image thresholding is divided into three sub tasks, which deal with the contrast image construction, the high contrast pixel detection, and the local threshold estimation, respectively. For high contrast pixel detection, Su et al. used Otsu's global thresholding to detect the desired high contrast image pixels which lay around the text stroke boundary. We implemented SIS thresholding [9] in place of Otsu's [10] in this step to achieve higher efficiency.

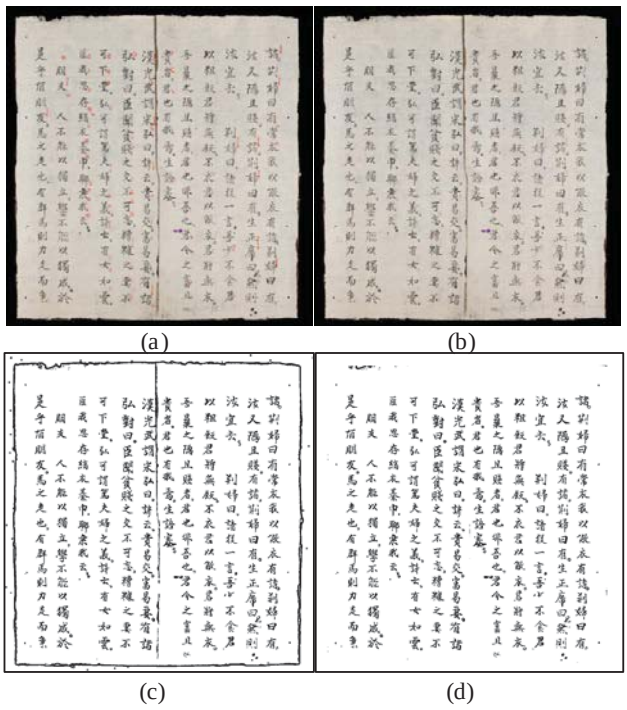


Figure 3. Results of preprocessing
(a) original image, (b) red comment removed image, (c) binarized image, (d) border line removed image

3.3 Border Line Removal

In most of Nom historical documents, a text region is often bounded by a frame as the left image of Figure 1(a) or an outline generated from the binarization process as the image of Figure 3(c). They may become noises that can affect the segmentation process. Therefore, we have improved the method using projection profile in [11] to remove marginal noises effectively.

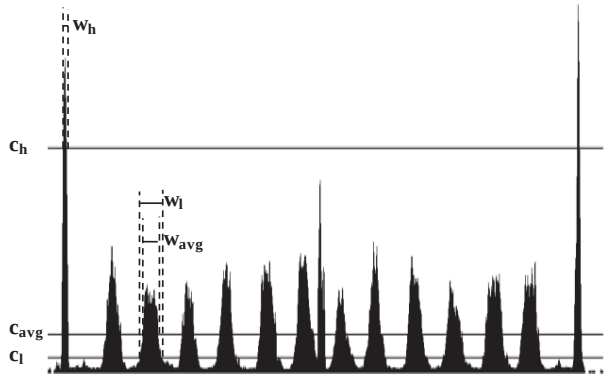


Figure 4. Vertical projection profile of Figure 4(c) and critical densities

We used three values called critical densities to detect border lines. These values are calculated from projection profile. Since average density c_{avg} is the first value, and the others are high and low critical densities come from the following equation:

$$c_h = \frac{\alpha_h c_{avg} + \beta_h d_{max}}{\alpha_h + \beta_h} \text{ and } c_l = \frac{\alpha_l c_{avg} + \beta_l d_{min}}{\alpha_l + \beta_l} \quad (1)$$

where d_{max} and d_{min} are the maximum density and minimum density. The parameters α, β are used to adjust the values of critical densities. Figure 4 shows critical densities of Figure 3(c) when $\alpha_h = \alpha_l = 1$, $\beta_h = 1.5$, and $\beta_l = 1.3$.

The border line detection analysis starts from left to right of projection profile to calculate its intervals w_h, w_{avg} and w_l when it is crossed by c_h, c_{avg} and c_l critical density lines, respectively. The zone will be detected as a border line if the width is narrower than others and almost has the same value in different positions on the profile. It is defined by the following condition:

$$w_{avg} < avg(w_h) \;\&\&\; \frac{w_{avg}}{w_h} < \gamma_h \text{ or} \quad (2)$$

$$w_l < avg(w_h) \;\&\&\; \frac{w_l}{w_{avg}} < \gamma_l \quad (3)$$

where $avg(w_h)$ refers to the average of intervals crossed by the high critical density c_h line on the profile. We set parameters γ_h and γ_l to 5 and 3 respectively. As illustrated in Figure 3(d), the border lines including boundaries and a separating line in the center are removed.

4. AREA VORONOI DIAGRAM

4.1 Area Voronoi Diagram

Let $G = \{g_1, \dots, g_n\}$ be a set of non-overlapping components in the two dimensional plane and let $d(p, g_i)$ be the Euclidean distance between a point p and a component g_i defined by:

$$d(p, g_i) = min_{q \in g_i}\, d(p, q) \quad (4)$$

Then the Voronoi region $V(g_i)$ and the area Voronoi diagram $V(G)$ are defined by:

$$V(g_i) = \{p | d(p, g_i) \le d(p, g_j), \forall j \ne i\} \quad (5)$$

$$V(G) = \{V(g_1), \dots, V(g_n)\} \quad (6)$$

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, and the points where Voronoi edges meet are called Voronoi points.

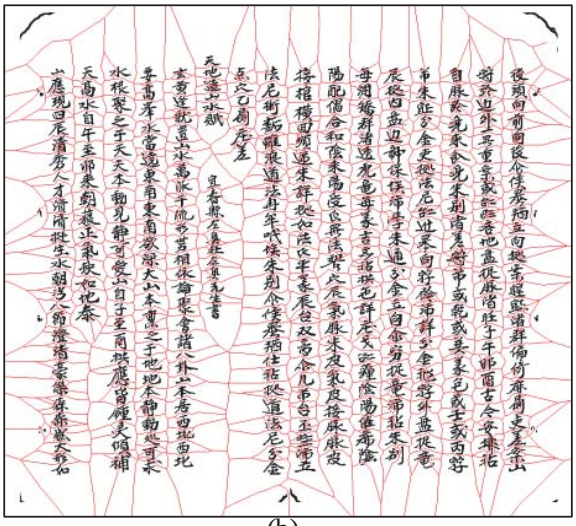
4.2 Construction of Area Voronoi Diagram

To construct the area Voronoi diagram, we utilize the approach presented in [6] with some modifications.

In labeling process, different with Kise et al. [6], we implemented the procedure based on contour tracing [12]. Moreover, in order to accelerate the construction of the area Voronoi diagram, connected components whose lengths of borders are less than or equal to T_n (filtering parameter) pixels are removed, and border points are sampled with the rate of T_s (sampling parameter) pixels. This process is called border point sampling. The sample points are then used to construct the area Voronoi diagram by the algorithm of the plane sweep method [13].

From the image in Figure 5(a), the area Voronoi diagram shown in Figure 5(b) is generated. As shown in this figure, the edges lie between characters or separate components of characters.

(a)



(b)

Figure 5. Construction of the area Voronoi diagram (a) binarized image and (b) area Voronoi diagram

5. SEGMENTATION

After the area Voronoi diagram is constructed, as shown in Figure 5(b) Voronoi edges lie between any adjacent connected components. Every character pattern seems to be represented as a set of Voronoi regions which are adjacent with one another. The process of character segmentation is thus considered to be grouping Voronoi regions, that is, deleting superfluous Voronoi edges which lie between connected components within a character pattern. In order to achieve this goal, we need criteria for selecting superfluous Voronoi edges from the area Voronoi diagram.

5.1 Criteria for Removing Voronoi edges

When the distance between 2 regions is narrower than those between columns or rows, the Voronoi edge between them can be deleted. Moreover, in order to restrict the grouped region to be not much larger than a character, so that we must estimate the size of characters and calculate the boundary size of a grouped component. Based on information of connected component labeling and the area Voronoi diagram, the standard features can be calculated as follows:

5.1.1 Relative distance

The distance between two bounding rectangles r_i and r_j of two connected components cc_i and cc_j which are easily obtained in labeling process, is calculated by:

$$d_r(i,j) = \begin{cases} -2, & r_i \subset r_j \\ -1, & r_j \subset r_i \\ d(r_i, r_j), & otherwise \end{cases} \quad (7)$$

5.1.2 Minimum distance

Let $E = \{l_1, \dots, l_n\}$ be a Voronoi edge between two connected components cc_i and cc_j, where l_k is a line segment of E. Each line segment l_k separates two points p_m and q_m on the borders of cc_i and cc_j, respectively. The minimum distance is then defined by:

$$d_e(i,j) = min_{1 \le m \le n}\, d(p_m, q_m) \quad (8)$$

5.1.3 Intersected area - $area_r(i,j)$

The area of rectangle intersected by two bounding rectangles of two connected components cc_i and cc_j.

5.1.4 Estimated size

Let $R = \{r_1, \dots, r_n\}$ be a set of rectangles bounding connected components, then estimated sizes are then assigned as the average width $\bar{w}$ and the average height $\bar{h}$ of rectangles in R.

5.1.5 Boundary size of grouped component

When grouping two components, the boundary is the union rectangle of those rectangles. The size (width $w_r(i,j)$ and height $h_r(i,j)$) of the grouped component is then simply the size of the union rectangle.

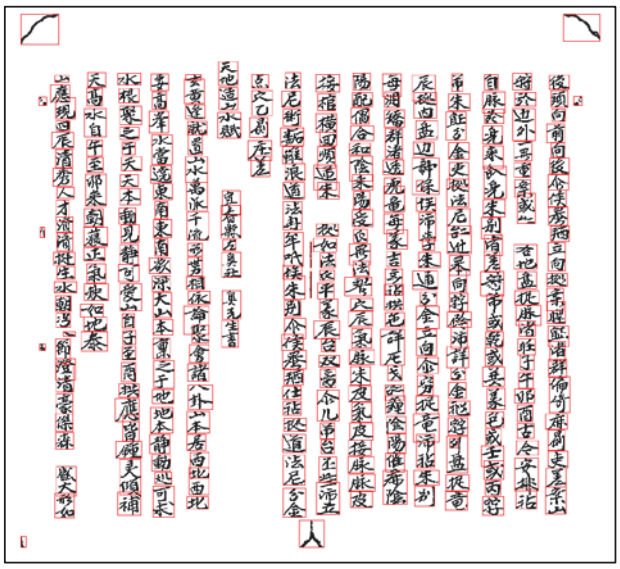
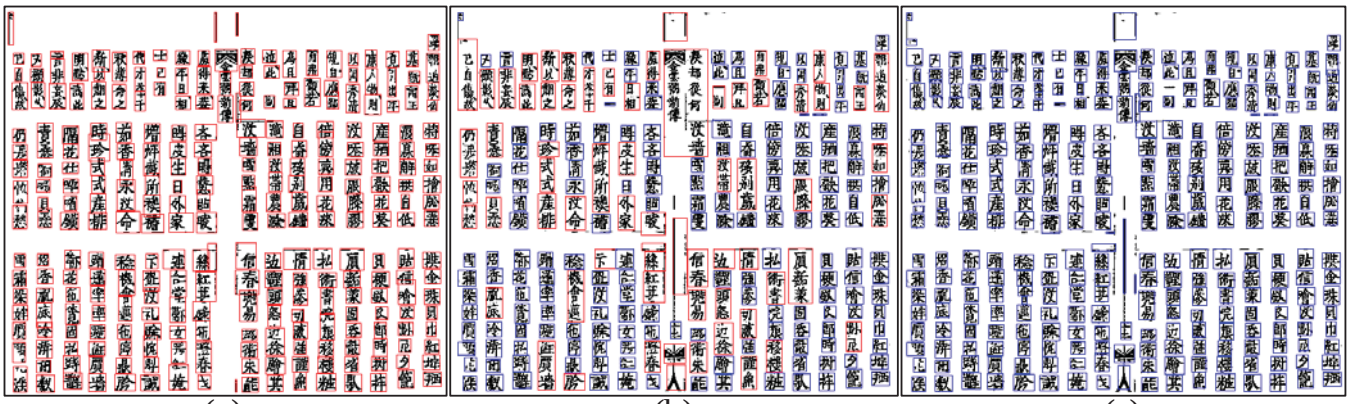


Figure 6. Result of character segmentation from Figure 5(a)

5.2 Segmentation Procedure

Voronoi edges are evaluated according to the above criteria. However, first of all, we generate the relations diagram between regions, namely we determine neighbors of each region and distances between them. Basically, if space between two adjacent components is too narrow with respect to the size, the border edge should be removed, and they are identified as components within a character pattern.



(a) (b) (c)

Figure 7. Segmentation results for the same page: (a) segmented characters by using area Voronoi diagram only, (b) split regions by RXYC method with $t_x = t_y = 0$ and (c) segmented characters by using area Voronoi diagram for every region partitioned via RXYC method

In order to evaluate all features in an efficient manner, we utilize the following criteria:

$$area_r(i,j) > \frac{min\big(area_r(i), area_r(j)\big)}{4} \;\&\&$$

$$\frac{w_r(i,j)}{\bar{w}} < T_{s1} \&\& \frac{h_r(i,j)}{\bar{h}} < T_{s1} \quad \text{or} \qquad (9)$$

$$d_e(i,j) < T_{d1} \;\&\&$$

$$\frac{w_r(i,j)}{\bar{w}} < T_{s2} \&\& \frac{h_r(i,j)}{\bar{h}} < T_{s2} \quad \text{or} \qquad (10)$$

$$(d_e(i,j) < T_{d2} \;||\; d_r(i,j) < T_{d2}) \;\&\&$$

$$\frac{w_r(i,j)}{\bar{w}} < T_{s3} \&\& \frac{h_r(i,j)}{\bar{h}} < T_{s3} \qquad (11)$$

where T_{d1} and T_{d2} are distance-related thresholds; T_{s1}, T_{s2} and T_{s3} are size-related thresholds. If two regions satisfy at least one of these equations, they are grouped.

With size-related thresholds, they are independent of resolution and layout. A grouped component cannot have size bigger than the character size, so that these thresholds are in range (1, 2). Moreover, based on ordering of priority, we should configure $T_{s1} < T_{s2} < T_{s3}$. Otherwise, with distance-related thresholds, they depend heavily on layout of documents. Thus, we use a frequency of distribution of minimum distance. The thresholds T_{d1} and T_{d2} are set to the average and the moment of this distribution.

5.3 Optimization via recursive X-Y cut

Since we recognize that in some documents, characters are arranged in rows and columns quite clearly. Moreover, the construction of the area Voronoi diagram is a time-consuming process. Therefore, we utilize the recursive x-y cut method to detect regions of character. The algorithm recursively splits the document into two or more smaller rectangular regions. At each step of the recursion, the vertical and horizontal projection profiles are computed and used to detect valleys. To compute the valleys n the projection profile histograms, noise removal thresholds t_x and t_y are used. The process continues until region has approximately the same size as a character or any valleys can be detect further. Which region has size larger than a character size, the area Voronoi diagram is then produced and characters in this region are segmented.

In the Figure 7, the result of recursive X-Y cut and segmentation results are showed.

6. EXPERIMENTS

The proposed method has been experimented on a dataset of 8 titles in different layouts and character sizes provided by the National Library of Vietnam. The original images are digitized at the resolution of 240 dpi. In order to accelerate the process, we reduced images to 120 dpi. From 525 images of these 8 titles, we randomly selected 5 images for each title, containing totally 12,905 characters for the experiment.

In this experiment, we compare the segmentation results in two methods, one utilizing the area Voronoi diagram only and the other optimized by using the recursive x-y cut method. In all pages, we set 5 and 7 to the sampling parameter T_s and the filtering parameter T_n and assigned 1.75, 1.45 and 1.25 to the size-related thresholds T_{s1}, T_{s2} and T_{s3}, respectively.

We evaluate the overall performance in term of F-measure. The F-measure is defined by the formula (12) where R is recall rate, P is precision rate and they are defined by the formula (13), (14) respectively.

$$F = \frac{2}{\frac{1}{R}+\frac{1}{P}} \qquad (12)$$

$$R = \frac{Number\ \ of\ \ correct\ \ segmented\ \ characters}{Number\ \ of\ characters\ \ in\ the\ target\ \ data} \qquad (13)$$

$$P = \frac{Number\ \ of\ \ correct\ \ segmented\ \ characters}{Number\ \ of\ segmented\ \ regions} \qquad (14)$$

The ground truth for these characters is defined and results are tested by hand, it means that the author verify by counting characters and segmented characters one by one. Table 1 summarizes the experimental results. The accuracy which is defined as the percentage of characters that are segmented correctly in F-measure has the value 80.19% and reaches to 85.77% in the improved version.

There are three sources of error: incorrect grouping, overlapped or touched characters and noisy regions. A considerable amount of errors come from pages containing characters of various sizes and curved text lines in their body text regions.

Table 1. Performance of Nom character segmentation

	Voronoi	x-y cut + Voronoi
Number of segmented regions	12,722	13,266
Number of correct characters	10,275	11,223
Recall (%)	79.62	86.97
Precision (%)	80.77	84.60
F-measure (%)	80.19	85.77

Table 2. Processing time for segmentation

	Voronoi	x-y cut + Voronoi
Number of segmented regions	12,722	13,266
Segmenting time (ms)	172,786	36,387
Average per region (ms)	13.58	2.74

In the improved version by x-y cut, a text region is divided into sub regions. The estimated character size of each region is then calculated in local, so that the size variation is reduced and the accuracy is improved. In fact, recursive x-y cut alone segments most of characters in neat layout of pages, and even if it leaves text regions without segmented, the area Voronoi diagram is constructed from each small region, so that the time complexity is reduced with higher accuracy. It is confirmed in Table 2 where the improved version is approximately 5 times quicker than the method employing area Voronoi alone.

7. CONCLUSION

This paper has presented an effective method for preprocessing and segmentation of Nom historical documents. We utilized simple noise removal and forcible binarization for preprocessing. Then we employed the area Voronoi diagram, which demonstrated its effectiveness in character segmentation. Information of neighborhood of connected components and relation between them were used to group components to form a character pattern. Moreover, to reduce size variation and decrease processing time for constructing the area Voronoi diagram, the recursive x-y cut method was employed effectively. The experimental results on a number of document images show that the proposed method is highly accurate. However, as said earlier, this research is just the beginning, there are many challenge from now. Further work will employ the Hough transformation to extract line segments before character segmentation, as well as develop OCR and utilize it in segmentation to improve the accuracy. The other work is to segment overlapping and touching character. Furthermore, we will use the OCR and the word-spotting technique to classify and label the segmented characters. Finally, we will experiment on a larger dataset to be able to evaluate the efficiency exactly. As a result, we construct a database of Nom character patterns.

8. ACKNOWLEDGMENTS

The authors thank the National Library of Vietnam and the Vietnamese Nom Preservation Foundation for providing Nom historical document pages. The authors also thank Mr. Su for helping us to implement the binarization function.

9. REFERENCES

[1] V.J. Shih, T.L. Chu, “The Han Nom Digital Library,”, in *The International Nom Conference*, The National Library of Vietnam, Hanoi, November 12-14, 2004.

[2] M.S. Kim, K.T. Cho, H.K. Kwag, J. H. Kim, "Segmentation of Handwritten Characters for Digitalizing Korean Historical Documents," *Document Analysis Systems* 2004, 114-124.

[3] L.Y. Tseng, R.C. Chen, "Segmenting handwritten Chinese characters based on heuristic merging of stroke bounding boxes and dynamic programming," *Pattern Recognition Letters* 19(10), 1998, 963-973.

[4] Y.H. Tseng, H.J. Lee, "Recognition-based handwritten Chinese character segmentation using a probabilistic Viterbi algorithm. *Pattern Recognition Letters* 20(8), 1999, 791-806.

[5] S. Zhao, Z. Chi, P. Shi, H. Yan, "Two-stage segmentation of unconstrained handwritten Chinese characters," *Pattern Recognition* 36(1), 2003, 145-156.

[6] K. Kise, A. Sato, M. Iwata, "Segmentation of page images using the area Voronoi diagram," *Comput. Vis. Image Underst.* 70(3), 1998, 370-382

[7] Y. Lu, Z. Wang, C.L Tan, "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.

[8] B. Su, S. Lu, C.L Tan, "Binarization of historical handwritten document images using local maximum and minimum filter," *International Workshop on Document Analysis Systems*, June 2010, 159–165

[9] J. Kittler, J. Illingworth, “Threshold selection based on a simple image statistics,” *Comput. Vision Graphics Image Process*.30, 1985, 125-147.

[10] N. Otsu, “A threshold selection method from gray-level histograms,” *IEEE Trans. System*, Man Cybernetics9, 1979, 62-66.

[11] W. Peerawit, A. Kawtrakul, “Marginal noise removal from document images using edge density,” In: *4th Information and Computer Engineering Postgraduate Workshop*, Phuket, Thailand, 2004.

[12] F. Chang, C. J. Chen, “A Fast Method for Labeling Connected Components in an image,” *IPPR Conference on Computer Vision, Graphics and Image Processing* (CVGIP), 2003, 327-333.

[13] A. Okabe, B. Boots, K. Sugihara, "Spatial Tessellations. Concepts and Applications of Voronoi Diagrams," J. Wiley and Sons, Chichester, 1992, 257-264.

Figure 8. Examples of input documents (the first image) and segmentation results of using area Voronoi diagram (the second image) and using RXYC to optimize (the third image)