Collecting Handwritten Nom Character Patterns from Historical Document Pages

Truyen Van Phan, Bilan Zhu and Masaki Nakagawa
Department of Computer and Information Sciences
Tokyo University of Agriculture and Technology
Tokyo, Japan
E-mail: truyenvanphan@gmail.com

***Abstract*—In this paper, we present methods of segmenting Nom historical documents and clustering character patterns to build a Nom character pattern database. Nom is an ideographic script to represent Vietnamese, used from the 10th century to 20th century. However, this heritage is nearly lost. In order to preserve the wisdom and knowledge expressed in Nom, recognition and digitalization are indispensable. Because there is no OCR for Nom yet, we have to start from collecting patterns. We have employed a projection profile based method for segmenting hundreds of pages into individual characters. Then, we have implemented a combination of Chinese OCR-based clustering and K-means clustering to group characters into categories. The experiment shows that the proposed system can help collecting the characters patterns effectively. Moreover, it has revealed that there are many character classes lost or uncategorized so far.**

Keywords-segmentation; clustering; Chu Nom; Han Nom; historical document; pattern collection; offline character database; document image analysis; Vietnamese ancient text.

I. INTRODUCTION

There exists a large amount of documents in Nom, an ideographic script, used in Vietnam for over 1,000 years since its independence in 939 AD. It makes use of Chinese characters to write Vietnamese, known as Hán Nôm (漢喃) in Vietnamese, as well as other characters coined following the Chinese model. From the 10th century and into the 20th, much of Vietnamese literature, philosophy, history, law, and so on were written in the Nom script. It is in danger of further destruction after more than 100 years of wars. Because the last national examination using Han Nom was in the 1919, scholars who master Han Nom are almost extinct. Only less than 100 scholars world-wide can read Nom now. Moreover, over 90% of documents have not been translated to the current Vietnamese yet. As a consequence, much of historical value of Vietnam is inaccessible to the 80 million speakers of the language [1].

It is extremely necessary to preserve and utilize this cultural heritage. Vietnamese agencies and world-wide foundations have been collecting thousand volumes of Han Nom textbooks [2]. Among them, the Han Nom Special Collection Digitization Project, a collaborative project between the National Library of Vietnam and The Vietnamese Nom Preservation Foundation [1], has scanned about 4,400 texts, and provided online access to over 1,300 texts with about 89,000 two-pages images as shown in Fig. 1.

(a) (b)

Figure 1. Examples of images for Nom historical documents

Due to the construction of digital library, valuable documents are being preserved and made available for public. However, there remains work to be done to make them really valuable heritage. They must be fully digitized, i.e., indexed, annotated, recognized, and hopefully translated into the current Vietnamese. Since the full digitalization is most difficult and time-consuming, with a large amount of documents and a decreasing number of experts, the process cannot be done in a short period of time by just manual typing method. Obviously, document recognition techniques can speed up the process.

Recently, the OCR-based techniques are often used for document segmentation and digitalization. The latest OCR techniques show high performances on modern printed materials. However, they are still limited to the historical documents. The difficulties come from problems such as degraded documents, complex layouts and blurred, damaged or partially lost characters. Furthermore, the main challenges with Nom are for both ancient and new characters with the ratio of 1:1 in the number of characters. The former are similar to traditional Chinese characters; however, there a lot of them are hardly used in modern texts. The latter are pure Nom characters which are invented by the Vietnamese. The number of Han Nom characters is over 30,000 characters now.

For these reasons, although most of archives are composed of handwritten Chinese characters, using Chinese OCR directly cannot show high performance. As a result, we have to build a specific OCR system for Nom with training patterns collected from Nom documents.

As the first step, we have developed a pattern collection system with two main processes: segmentation

978-0-7695-4661-2/12 $26.00 © 2012 IEEE
DOI 10.1109/DAS.2012.25

and clustering. As most of documents have simple layout and nearly identical character size, we have implemented a projection profile based method in the segmentation step. Then OCR module is used to classify each of segmented characters and K-means is employed to cluster the rejected characters from recognition results. Therefore, this system is designed for the combination of both automatic and manual operation to specify labels for characters.

The remainder of this paper is organized as follows: the proposed system is overviewed in Section II; the preprocessing and segmentation process is described in Section III and Section IV, respectively; the details of the clustering method is presented in section V; some experiment results are showed in Section VI; and our concluding remarks are given in Section VI.

II. System Overview

The proposed system for collecting Nom character patterns from document images consists of three main steps: 1) preprocessing, 2) segmentation, 3) clustering by recognition and clustering by K-means. Overview of the system is drawn in Fig. 2.

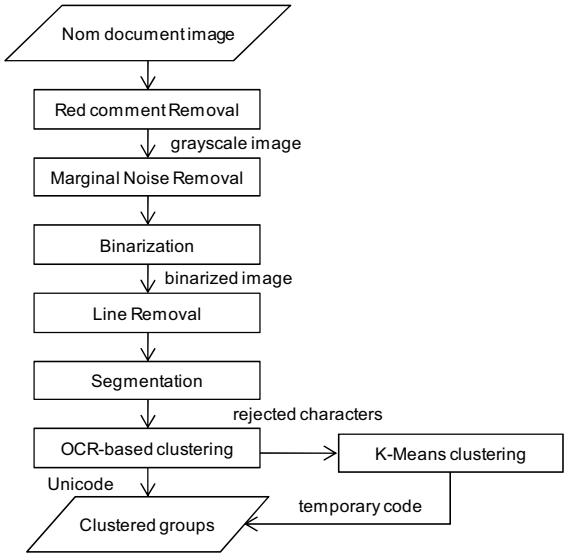


Figure 2. Overview of the system

In preprocessing, we firstly turn the red color of comments to light gray color to remove them after binarization. After the marginal noises of image are removed, the image is binarized by an effective method for badly degraded historical documents [3]. Finally, the lines which cover the text region or separate two pages are removed.

In the next step, to segment document images into individual characters, we apply the recursive X-Y cut (RXYC). The input document image is separated into several character regions and the regions are split into primitive segments recursively. Each primitive segment should be an individual character or part of a character.

After all, a combination of OCR-based clustering and K-means is used to classify the category of segmented characters.

III. Preprocessing

A. Red Comment Removal

In most of documents, beside characters there exist many red comments that affect binarization. It is necessary to remove them. In this system, we use the HSL filter to detect the red color of comments. We specified the range of hue, saturation, luminance as (300, 50), (0.3, 1.0), (0.3, 1) respectively. If pixel's color is inside of this specified range, it is turned to light gray which will be removed by binarization.

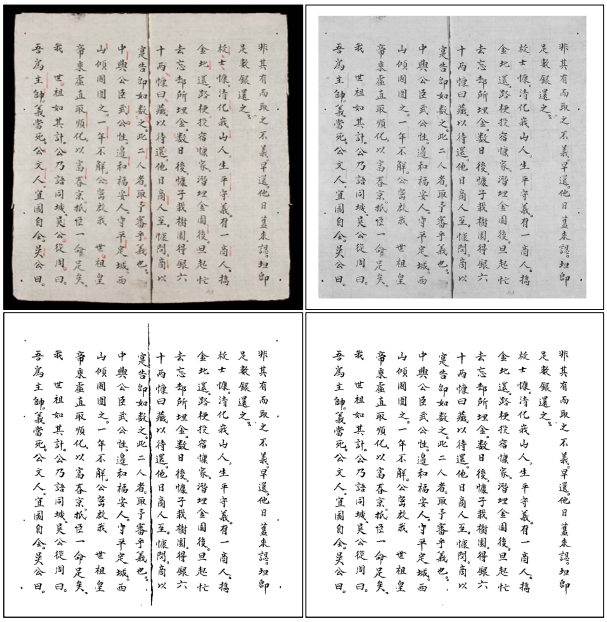


Figure 3. Results of preprocessing: (a) original image, (b) red comment and marginal noise removed image, (c) binarized image, (d) line removed image

B. Marginal Noise Removal

In digitized books, marginal noises often appear along the page border. In order to remove them, we use the projection profile analysis. Projection profiles are calculated on the grayscale image. Then the new border of image is determined by detected margin values as shown in Fig. 4.

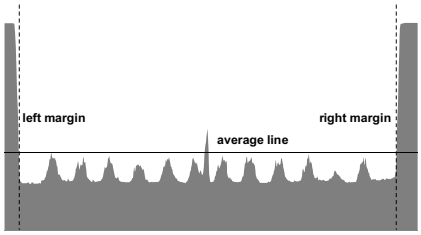


Figure 4. Margin detection on horizontal projection profile of Fig. 3(a)

C. Binarization

Since most historical archive document images are of poor quality due to aging and ink fading, binarization has a significant impact on the OCR performance. Among them,

we utilize the method proposed by Su et al. [3] because it has been proved effective in Document Image Binarization Contest (DIBCO) 2009 [4]. 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. 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 employ SIS thresholding [5] in place of Otsu's [6] in this step to achieve higher efficiency.

D. Line Removal

In Nom historical documents, a text region is often bounded by a frame as the image Fig. 1(b) or an edge between two pages as the image Fig. 3(c). We use three values called critical densities to detect lines. These values are calculated from projection profile. The average density c_A is the average among all the projections. High and low critical densities c_H and c_L are the weighted averages between the average density c_A with the maximum and minimum densities, respectively.

The line detection analysis starts from left to right to calculate intervals w_H, w_A and w_L when it is crossed by c_H, c_A and c_L critical density lines, respectively on the projection profile. A line is detected if w_H, w_A and w_L are small and had nearly the same values.

IV. SEGMENTATION

The segmentation procedure is based on the recursive X-Y cut method. The document is split into two or more smaller rectangular regions, recursively. At each step of the recursion, the horizontal and vertical projection profiles of each region are computed. Then the bottoms between valleys are detected by using thresholds, and smaller regions are divided by these bottoms. The process continues until the size of region approximates to the estimated character size.

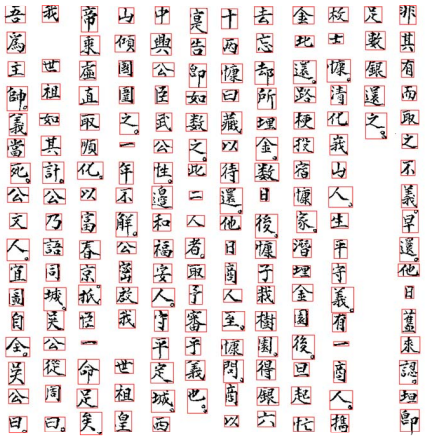


Figure 5. Result of character segmentation from Fig. 3(d)

The estimated character size is calculated based on the average of intervals crossed by the average density line c_A and the low critical density line c_L.

We run the recursive segmentation two times. In the first time, the recursion is used to segment a document into lines and separated characters. The bottom detected threshold is initialized with value of 0, then increased 1.5 after each loop. And in the second time, the recursive segmentation is used to segment overlapping or touching characters. The bottom is detected with the minimum value between two valleys in the projection profile.

V. CLUSTERING

After the segmentations process, hundreds of pages are segmented into individual characters. In the clustering process, we firstly use an OCR to recognize these characters. Because the segmented characters shapes are quite different from the trained character shapes, we have specified a threshold for each class determined experimentally. If the recognition score are smaller than this threshold values, the character is identified with the recognition code, if not it is rejected. Then, we use K-means to cluster rejected characters into groups.

A. OCR-based Clustering

In Nom historical documents, there are about 50% characters borrowed from Chinese. We apply offline Chinese character recognition to specify the code of segmented characters and cluster them into groups. The recognition system consists of three steps: nonlinear normalization, feature extraction and classification.

A segmented character is normalized using line density projection interpolation (LDPI) [7]. After normalization, we apply the gradient feature extraction. The gradient vector $g(x,y) = [g_x, g_y]^T$ at a pixel (x,y) in a normalized image is computed by using the Sobel operator as indicated in the following formulas:

$$\begin{aligned} g_x(x,y) &= f(x+1,y-1) + 2f(x+1,y) + f(x+1,y+1) \\ &\quad -f(x-1,y-1) - 2f(x-1,y) - f(x-1,y+1) \\ g_y(x,y) &= f(x-1,y+1) + 2f(x,y+1) + f(x+1,y+1) \\ &\quad -f(x-1,y-1) - 2f(x,y-1) - f(x+1,y-1) \end{aligned}$$

We use eight direction planes, corresponding to eight chain code directions (Fig. 6(a)). Each gradient vector is decomposed into two components in two neighboring chain code directions (Fig. 6(b)).

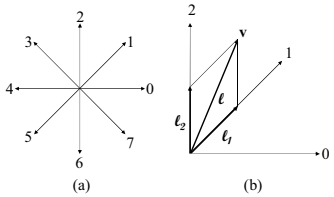


Figure 6. Eight chain code directions (a) and directional decomposition of a gradient vector (b)

The size of normalized plane and direction plane is set to 48x48 pixels. From each of 8 direction planes, we extract 8x8 feature values, as a result, obtain totally 512 features. For reducing the classifier complexity and improving classification accuracy, we reduce the dimensions from 512 to 160 by Fisher linear discriminant analysis (FLDA) [8].

For classification, we use two classifiers: the Euclidean distance classifier and MQDF2 [9]. We select 100 candidate classes according to the Euclidean distance. MQDF2 is then computed on the candidate classes only. We use 50 principal eigenvectors for each class.

After the segmented character is recognized, we get a list of candidates. In ascending order, we check the score of every candidate, and accept the character with candidate code if the candidate score is smaller than the class threshold, otherwise we reject it. The process is illustrated in Fig. 7.

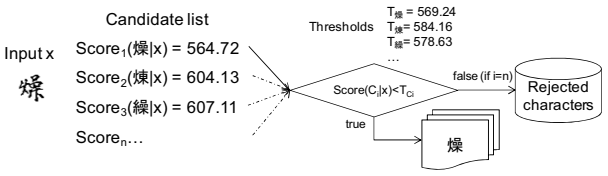


Figure 7. Clustering and rejection based on recognition score

The class threshold is calculated based on the recognition score when we test the recognizer after training.

B. K-means Clustering

There are many rare Chinese characters and new Nom characters that they are rejected from the recognition-based clustering. In order to reduce cost in manual verification and typing, we use K-means clustering to group these characters. Then the characters in a group are shown to the operator to verify the correctness of the grouping and input the characters code of the group.

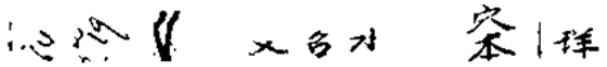
K-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. The algorithm works by first selecting k locations at random to be the initial centroids for the clusters. Each observation is then assigned to the cluster which has the nearest centroid, and the centroids are recalculated using the mean value of assigned values. The algorithm then repeats this process until the cluster centroids do not change anymore. We employ the standard algorithm of K-means which is referred as Lloyd's algorithm.

VI. EXPERIMENT

The proposed method has been experimented on a dataset of 5 titles with 414 page images 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 reduce the images to 120 dpi.

The OCR used in this system was trained on 4 large databases: ETL9B, JEITA-HP (DATASET-A and DATASET-B), NTT-AT, Kuchibue and Nakayosi [10]. The first 2 databases are offline character pattern database while the rest 3 databases are converted from online patterns. The total numbers of categories and characters of the training set are 4,060 and 3,983,823 respectively. Since our aim is not to evaluate our recognizer, we have used the entire set to train the recognizer and calculate the score threshold for each category which is used in the clustering procedure. The recognition rate is 99.24%. Because the quality of the database characters and the document characters are quite different, we add an extra value $\alpha = 200$ to each score threshold.

In this experiment, we evaluate the performance of segmentation and clustering on each title. In segmentation, because there do not exist the ground truths of the images, the accuracy is defined as the percentage of correct segments and segmented regions. We define three types of errors in segmentation: noise regions, fragments and compounds as Fig. 8. Noise regions are tears, ink stains or comments, etc. Fragments are only one or multiple components of an individual character pattern. Compounds are one or more than one complete character patterns with others.



a) noise regions b) incorrect segments c) over-segments

Figure 8. Examples of some errors in segmentation

Table 1 summarizes the experimental results. As the segmentation algorithm is performed, separated from recognition, obviously most of errors are fragments or compounds. These errors are caused by the insufficiently estimated size, improper criteria to segment characters by projection profiles. However, the accuracy of approximately 90% is obtained at this time.

TABLE I. SEGMENTATION RESULTS

	No. of segmented regions	No. of noise regions	No. of fragments	No. of compounds	No. of correct segments	Accuracy (%)
Title 1	13,482	627	372	188	12,295	91.20
Title 2	37,685	116	292	740	36,537	96.95
Title 3	13,696	58	1,286	576	11,776	85.98
Title 4	12,947	244	934	1,020	10,739	82.95
Title 5	19,687	141	1,669	1,616	16,261	82.60
Total	97,497	1,186	4,553	4,140	87,608	89.86

In clustering evaluation, error-removed character patterns with categorized labels were prepared. Then, the character patterns were recognized by OCR. Based on the threshold of each class, a character is accepted and collected into a group as Fig. 9 or rejected and clustered by K-means as Fig. 10. The parameter k in K-means is set to 500.

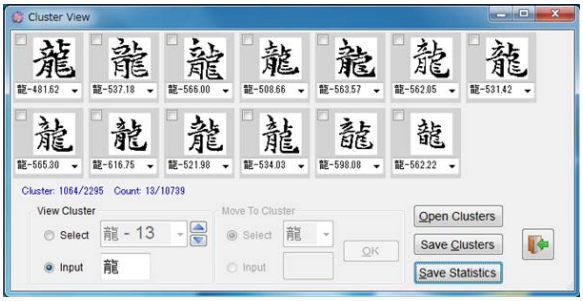


Figure 9. An example of recognition-based clustering result

TABLE II. EXPERIMENT RESULTS OF RECOGNITION-BASED CLUSTERING AND K-MEANS CLUSTERING

		Recognition-based clustering				K-Means clustering			Overall	
	No. of segments	No. of recognized classes	No. of accepted characters	No. of correct recognized characters	Cluster accuracy (%)	No. of rejected characters	No. of correct clustered characters	Cluster accuracy (%)	No. of correct clustered characters	Cluster accuracy (%)
Title 1	12,295	2,117	11,333	8,031	70.86	962	798	82.95	8,829	71.81
Title 2	36,537	2,712	33,994	26,377	77.59	2,543	1,964	77.23	28,341	77.57
Title 3	11,776	1,953	10,286	7,359	71.54	1,490	1,107	74.30	8,466	71.89
Title 4	10,739	1,795	9,931	8,043	80.99	808	720	89.11	8,763	81.60
Title 5	16,261	2,437	12,587	6,193	49.20	3,674	2,045	55.66	8,238	50.66
Total	87,608	3, 403	78,131	56,003	**71.68**	9,477	6,634	**70.00**	62,637	**71.50**

The accuracy of the recognition-based clustering is the percent of the number of the correct recognized characters and the number of the accepted characters. On the other hand, the accuracy of the K-means clustering is the percentage of majority characters among all. Table 2 shows the results of the clustering process.

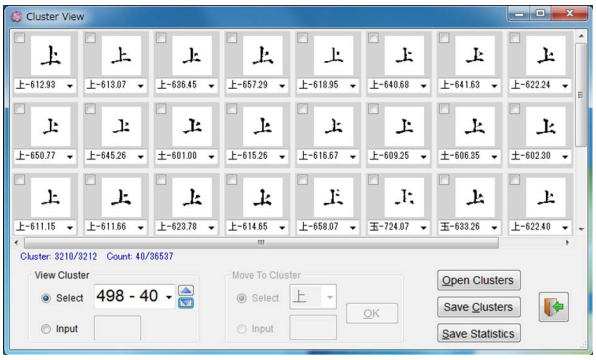


Figure 10. An example of K-means clustering result

Although some characters are well-classified by K-means, they are rejected by insufficient thresholds in the recognition-based clustering step, as shown in Fig. 10.

Moreover, from the results, we see that the greatest part of incorrect recognized characters are new or rare characters which have similar shapes with categorized characters in dictionary as shown in Fig. 11. On the other hand, it has revealed that there are many character classes lost or uncategorized so far as shown in Fig. 12.

Figure 11. Characters misclassified into "妖" group

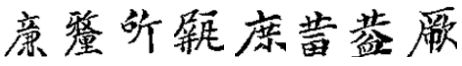


Figure 12. Some rare and uncategorized characters

VII. CONCLUDING REMARKS

This paper described our attempts to collect Nom character patterns from historical document pages. A projection profile-based method is employed to segment document images into individual characters. Then a recognition-based method is used to clustered characters into specified categories. After all, the rejected characters are grouped into unspecified categories. The images in a group are shown to an operator to verify the correctness of the grouping. This work is desirable for saving time and effort to collect character patterns now and fully digitalize documents in the future. In order to improve the accuracy of the recognition and clustering, we intend to build a bigger dictionary by generating variations artificially from Nom or traditional Chinese fonts. Optimizing threshold in recognition-based clustering and implementing a better clustering than K-means are also our future works.

ACKNOWLEDGMENT

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.

REFERENCES

[1] http://nom.nlv.gov.vn/nlvnpf/index.php

[2] 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.

[3] 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.

[4] B. Gatos, K. Ntirogiannis and I. Pratikakis, "ICDAR 2009 Document Image Binarization Contest (DIBCO2009)", 10th International Conference on Document analysis and Recognition (ICDAR'09), July 2009, 1375-1382.

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

[6] N. Otsu, "A threshold selection method from gray-level histograms," IEEE Trans. System, Man Cybernetics 9, 1979, 62-66.

[7] C. L. Liu, and K. Marukawa, "Pseudo two-dimensional shape normalization methods for handwritten Chinese character recognition," Pattern Recognition, vol. 38, no. 12, Dec 2005, 2242-2255.

[8] K. Fukunaga, Introduction to Statistical Pattern Recognition, 2nd edition, Academic Press, 1990.

[9] F. Kimura, K. Takashina, S. Tsuruoka, and Y. Miyake, "Modified quadratic discriminant functions and the application to Chinese character recognition," IEEE Trans. PAMI, vol. 9, no. 1, 1987, 149-153.

[10] M. Nakagawa, and K. Matsumoto, "Collection of on-line handwritten Japanese character pattern databases and their analysis," Int. J. Document Anal. Recognit. 7(1), 2004, 69–81.