Digitalizing Scheme of Handwritten Hanja Historical Documents

Min-Soo Kim, Man-Dae Jang, Hyun-Il Choi, Taik-Heon Rhee and Jin-Hyung Kim
Korea Advanced Institute of Science and Technology
Artificial Intelligence and Pattern Recognition Laboratory
373-1, Guseong-dong, Yuseong-gu, Daejeon, 305-701, Republic of Korea
{mskim, mdjang, hichoi, three, jkim}@ai.kaist.ac.kr

Hee-Kue Kwag
AI R&D center, Dongbang SnC Co.,Ltd.
10th Floor, BaekSang Bldg, Gwanhun-dong, Jongno-gu, Seoul, Korea
hkkwag@dbsnc.co.kr

Abstract

In Korea, the historical archives of Lee dynasty have been digitized for years. Although the metallic pressing technology had been used in Korea as early as 13th century, most of these documents are written by the King's chroniclers and secretaries. In addition, ancient characters which are not used in contemporary texts take considerable proportion. As a consequence, it is extremely difficult to utilize conventional OCR systems, and most of the process has been performed manually. However, this manual processing has been unsatisfactory in terms of costs and efficiency. As an alternative, we built a system that is composed of a dedicated handwritten Hanja recognizer and an easy-to-use verifier. Preliminary experiments show that the proposed system can help enhancing the overall efficiency of the process and reducing the costs.

1. Introduction

The historical documents are obviously invaluable. From various kinds of documents such as record of historical events, description of ancient culture, a king's dairy and classical literature, we could learn a lot of things about the past. As time goes on, we are supposed to set much value on the historical documents.

Korean national agency also has been preserving historical documents from the past. However, just keeping the documents physically has limitation in maintenance and accessibility. For the maintenance, since the documents were written on papers, the government spends large amount of budget to keep the documents safely. The maintenance causes limited access to those documents, i.e., accessibility. Although many peoples want to read valuable documents, it is restricted to those who are authorized to do so for security.

The construction of digital library for historical documents may reduce maintenance cost, expand accessibility. It is easy to perform data management and information retrieval efficiently using digital archives if the documents were archived digitally. There are no more limitations in maintenance and accessibility. As the digitalization is expected to enlarge utilization historical documents efficiently, the Korean governments have been digitalizing historical documents from several years ago.

The digitalization of the documents related to science technology, education, culture and history is active in Korea. For instance, National Computerization Agency started digitalization of some historical documents from 2000. The project will be continued to 2005. However, the project covers only 5% of historical documents, i.e., 95% of documents are still remaining to be digitalized (Table 1). The reason why it takes too much time for digitalization of historical documents is explained in the following section.

Kind of Literatures	Total sum (unit:volume)
Documents	2,365,561
Books	2,034,871
Others	5,868,689

Table 1. Statistics for the number of classical literatures

From 108 B.C., The Korean began using Chinese symbols and characters in writing as well as learning about Chi-

Proceedings of the First International Workshop on Document Image Analysis for Libraries (DIAL'04)
0-7695-2088-X/04 $20.00 © 2004 **IEEE**

nese ideological systems until development of Hangul by The Great King Se-jong at 1443. Although Chinese symbols and characters were used in the period, those were slightly different from symbols and characters of Chinese in terms of meaning, shape, and usage statistics. We call such characters and symbols *Hanja*. The following figures show examples of historical documents written in Hanja.

Figure 1. Examples of image for Hanja historical documents

However, since Hanja basically came from Chinese, most characters seem to be similar and still used in writing, but not so much as Hangul.

The digitalization of such documents has several important steps to enrich utilization. The first one is translation Hanja into Hangul. The translation is most difficult and time-consuming step of digitalization. Since the native language of Korea is Hangul and Hanja is rarely used in modern documents, some experts in Hanja are needed to do translation[1]. Moreover, the contents of documents are quite special for certain fields, for example, diaries of a King, classical literatures, etc. Therefore highly educated experts are supposed to give more accurate translation than normal people's one. The second one is that there are so many documents to be digitalized. With limited experts, digitalization process could be slowed down because it is very laborious.

Translation of large amount Hanja documents can be tackled by two approaches: *manual typing* and *use of OCR*. In manual typing, a Hanja character is divided into several strokes. All strokes in Hanja have different codes so that combination of stroke codes makes one character code. Although it seems to be efficient, a Hanja needs several typing strokes and there are so many character classes. That means it takes time for an operator to learn typing method such as stroke codes, combination rule, etc. Consequently the whole digitalization process would be blocked.

The use of OCR technique has been getting attention because the latest OCR technique shows high performance on modern printed materials. But digitalization of Hanja documents is not so easy with just OCR technique. Its primary difficulty comes from shape variation due to writers' habits, styles, and so on (refer to Figure 2(a)). Also, a blurred text is often appeared in the documents because of the complex structure of its strokes and their ink fading (refer to Figure 2(b)). According to the facts deteriorating the performance of character recognition, it is impossible to expect the perfect output of OCR, and consequently, it cannot simply substitute the manual typing. Also the utility of OCR to historical documents is very restricted[1][4][6][9].

Figure 2. Sample Hanja images

In this paper, we suggest an OCR-based technique to speed up the digitalization process. Proposed method is a combination of both typing and OCR to compensate one's drawback by others. From scanned page of documents, segmentation is performed and each segmented character image is fed to Mahalanobis distance based classifier to get a label. Segmentation and classification is repeatedly performed until all documents are processed. Once classification is done, we group characters that seem to be a same class with certain confidence and show them to an operator to verify the result (Fig 3). Whenever the operator finds misclassified character, he or she sets aside the character (Fig 4) to rejected class. When the verification is over, a label of the group is automatically assigned by the system.

The proposed method has several advantages in our framework. 1) *Fast labeling with high accuracy*; by glancing the group, an operator has ability to input a label a total of correctly classified characters. 2) *Cost effective in typing*: we can control reliability of classification by introducing rejection system. When we compared full manual typing cost and reforming cost, reforming cost was much higher than that of manual typing. To our knowledge, rejection system reduced total input cost. 3) *Relatively easy training of*

1 Hangul has different grammar, structure and pronunciation compared to Hanja.

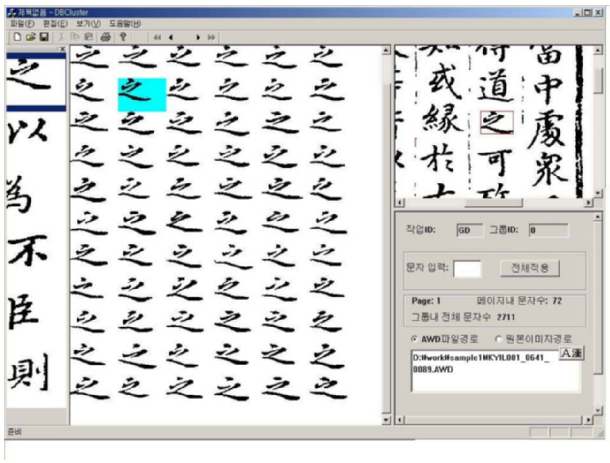

Figure 3. The example of characters with the same class label

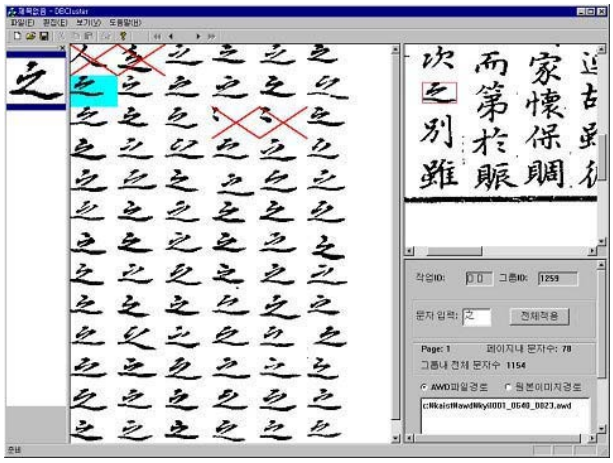

Figure 4. The example of elimination of misclassified characters by operator

classifier; there are numerous classes while we have limited number of data, we expected neural network (NN) family such as Support Vector Machine and Neural Networks doesn't seem to be well trained. Even thought NN family has been trained, it is likely to be biased.

From the experimental result on a Hanja handwritten document, we believe that the digitalizing process can be preformed with high quality and high-speed with the proposed system. The paper is organized as follows: the proposed digitalizing system is overviewed in Section 2; the details of. Construction of handwritten Hanja recognition and rejection system is described in Section 3; some experimental results are shown in Section 4; and our conclusive remarks are given in Section 5.

2. Explanation of proposed digitalizing paradigm

The proposed Hanja historical documents digitalizing scheme consists of three main modules: 1) preprocessing and segmentation, 2) Hanja recognition and rejection with OCR, 3) correction, data manual typing and verification, as shown in Figure 5.

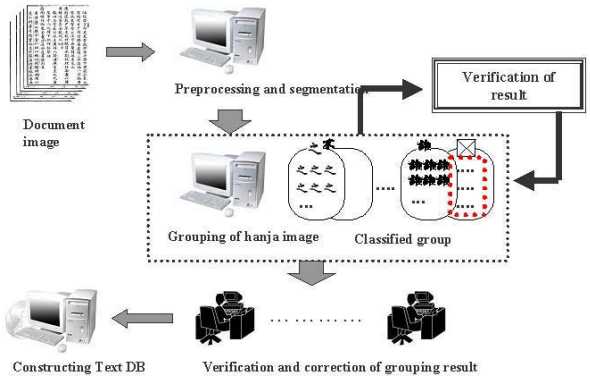

Figure 5. The proposed digitalizing system diagram

In the preprocessing and segmentation step, several hundred pages are segmented into individual characters rapidly. Then the OCR module is called up to identify the character class for each of the segmented characters. The handwritten OCR module consists of nonlinear shape normalization, extraction of meshed contour directional feature, and recognizer base on Mahalanobis distance (LDA). After the identifications are made, the characters with the same class label are collected into a predefined group. The images in a group are shown to the operator to verify the correctness of the grouping. Those images that do not actually belong to the class are deleted by a graphical user interface scheme. Guaranteeing correctness of the grouping, the characters code of the group is finally inputted. There needs only one typing for a group. It saves lots of laborious typing effort when it compared with the scheme in that one typing for each character when it appears.

3. Construction of handwritten Hanja recognition and rejection system

In Hanja grouping module, handwritten character recognition and rejection system consists of three steps: nonlin-

ear shape normalization, extraction of meshed contour directional feature, and classification as shown in Figure 6.

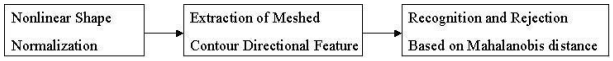

Figure 6. Handwritten Hanja Recognition and Rejection system

3.1. Nonlinear shape normalization

In order to compensate for shape distortions in handwritten characters, a Hanja image is first transformed nonlinearly into normalized one, 64×64 pixels. And then, we can obtain 8×8 blocks by putting 8×8 pixels into a block in both the horizontal and vertical directions. In previous works with the shape normalization[5][8], we take three methods based on dot density, crossing count, and line interval, respectively (refer to Figure 7). From the preliminary examination with the methods, we determine which one is more adequate to our data.

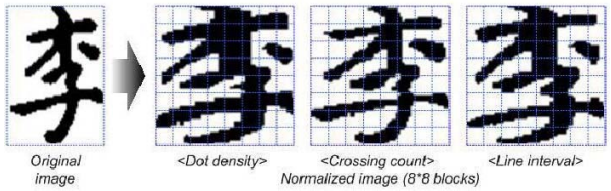

Figure 7. Normalized images with 8×8 blocks

3.2. Contour direction feature extraction

After the normalization of an original image, we extract the contour direction features representing the number of contour pixels in four main directions: horizontal, vertical, diagonal and inverse diagonal. A 3×3 window is defined over each contour pixel, as shown in Figure 8, and the pixel orientation is computed by its gradient magnitude[1][6].

Let $a(x, y)$ represent the stroke-direction angle at the black pixel(x,y) with respect to the x-axis. We define $a(x, y) = \tan^{-1}(G_x/G_y)$, where G_x and G_y are derivatives on the x-axis and y-axis, respectively. Based on the Sobel operator, the derivatives are defined as $G_x = (z_6 + 2z_7 + z_8) - (z_1 + 2z_2 + z_3)$ and $G_y = (z_3 + 2z_5 + z_8) - (z_1 + 2z_4 + z_6)$. The ranges

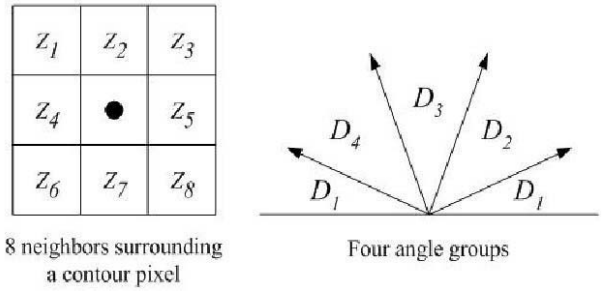

Figure 8. 3×3 window and four angle groups

of stroke angles are from $0\,^\circ$ to $180\,^\circ$. Since a Hanja usually consists of the strokes with the above directions, the angle range is partitioned equally into four groups, D_1, D_2, D_3 and D_4, as shown in Figure 6. Thus, we count the number of contour pixels belonging to each angle group, and extract a 4-dimension directional feature from each block. Consequently, we can obtain a 256-dimension $(8 \times 8 \times 4)$ feature vector.

The direction codes of stroke contour have long been used in character recognition and have been proved to be efficient to improve the recognition performance[1][2][3][6][7].

However, in some cases the local direction of contour pixel does not represent the genuine direction of stroke, especially when the size normalization produces staircases in character contour. Figure 9 shows a contour extracted from a Hanja image. In the figure (a), we can intuitively observe that Hanja mainly consists of the strokes with diagonal and inverse diagonal directions, but, the component rations of contour pixels in four directions are 22% for horizontal, 28% for vertical, 22% for diagonal, and 28% for inverse diagonal, respectively. As the result, the contour directional feature with the staircase problem cannot well represent the statistical property of stroke directions.

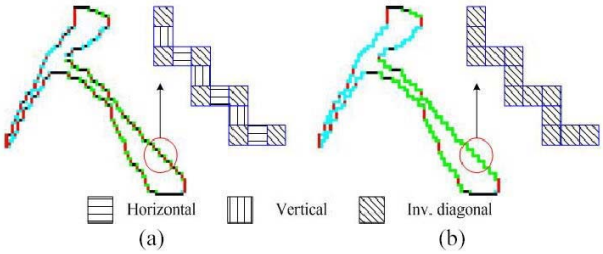

Figure 9. The contour directions

To overcome the problem, we produce the analog plane

for normalized digital image, and compute the pixel orientation on it. The simplest way to produce the analog plane is in essence weighted averaging a digital plane locally. It is perhaps easier to see the process as placing 3×3 mask over the pixel and summing the product of the mask value and the pixel value (refer to Figure 10).

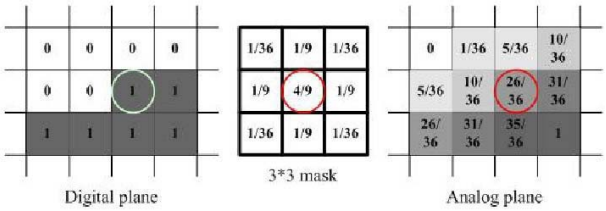

Figure 10. The analog plane using the 3×3 mask

After that, the staircases in contour direction will disappear or be compressed, and the local direction of contour points will accord with the stroke direction (refer to Figure 9 (b)). Also, the component rations of contour pixels in four directions are 11% for horizontal, 14% for vertical, 32% for diagonal, and 43% for inverse diagonal, respectively.

3.3. Recognition and rejection based on Mahalanobis distance

The classification rule is identical to the one that maximizes the posterior probability. We would allocate new observation $\mathbf{x}$ to that population ω_i, for which $P(\omega_i|\mathbf{x})$ is largest. Namely, we classify $\mathbf{x}$ to ω_i if $P(\omega_i|\mathbf{x}) > P(\omega_j|\mathbf{x})$ for all $i \neq j$.

$$P(\omega_j|\mathbf{x}) = \frac{P(\mathbf{x}|\omega_j)P(\omega_j)}{P(\mathbf{x})}, P(\mathbf{x}) = \sum_{k=1}^{g} P(\mathbf{x}|\omega_k)P(\omega_k)$$

Also, if we assume $\mathbf{x}|\omega_i$ represent a random samples form multivariate normal distribution with mean vector $\mu_\mathbf{i}$ and covariance matrix $\mathbf{\Sigma}$. Namely, under the assumption

$$P(\omega_j|\mathbf{x}) = \frac{\exp(-.5 \times r_j^2)}{\sum_{k=1}^{g} \exp(-.5 \times r_k^2)}$$

$$, where\ r_j = \sqrt{(\mathbf{x} - \mu_\mathbf{j})' \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu_\mathbf{j})}$$

Also, we reject $\mathbf{x}$ (classify $\mathbf{x}$ as rejection class) in case that maximum posterior probability is less than threshold value θ, determined experimentally. The following figure shows process of rejection system.

In Hanja classification with OCR, a very important problem is how many character classes are chosen as a relevant

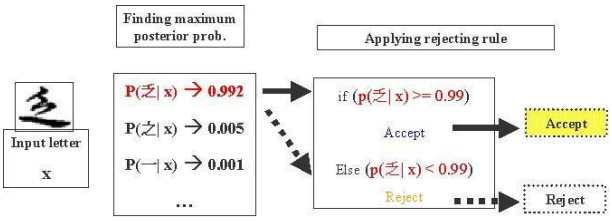

Figure 11. An example of applying rejection rule

set in recognition. All Hanja classes don't need to be necessary in the recognition of documents, because many characters rarely appear in common use. and this greatly increases the computational complexity of the recognition.

We have statistically investigated the ancient Korean documents about 3,896 pages containing about 1.5 million characters over 5,599 Hanja classes of Seungjungwon-Diary vol. 29. From the investigation, we observed that about 2,000 Hanja classes were frequently used, but about 3,600 Hanja classes were rarely used. Thus, we determined that 2,568 Hanja classes, which frequently appear about 99% in the documents, should be considered in the recognition step.

4. Experiments

In this section, we will present the results of a recognition experiment in which we evaluated with characters segmented from Seungjungwon-Diary vol. 29, a Korean government document, written by many writers during nearly 500 years. Figure 12 shows a document image and segmented characters of Seungjungwon-Diary vol. 29.

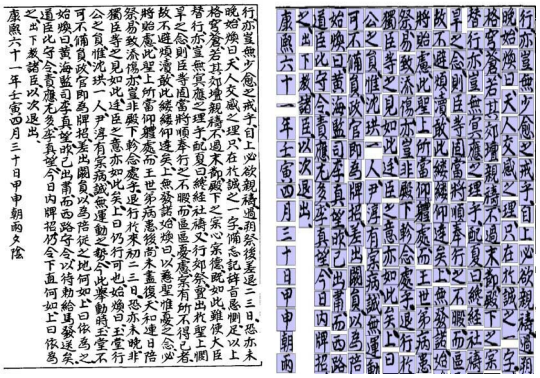

Figure 12. An example of segmented characters in Seungjungwon Diary vol.29 document

Proceedings of the First International Workshop on Document Image Analysis for Libraries (DIAL'04)
0-7695-2088-X/04 $20.00 © 2004 **IEEE**

Experiment 1: classification only

To show the performance of the classifier based on Mahalanobis distances, we carried out an experiment to compare the recognition rates of the classifiers based on Mahalanobis distance and Euclidean distance. The 100 characters per class extracted from Seungjungwon-Diary vol. 29 were used for training. Also, total 1,000 pages (divided into 5 parts, 200 pages) extracted from Seungjungwon-Diary vol. 29 were used for testing. E-classifier and M-classifier mean classifiers based on Euclidean distance and Mahalanobis distance, respectively.

The illustrations shown in Table 2 are summarized as follows. On the whole, the recognition rates of the M-classifier are superior to those of E-classifier. Also, we can find that total average recognition rate of M-classifier is more 4.6 percent point high than that of E-classifier.

Recognition rates	E-classifier	M-classifier
Set1	70.8	75.0
Set2	83.7	89.1
Set3	86.0	90.0
Set4	86.0	90.5
Set5	89.6	94.6
Total	83.2	87.8

Table 2. Comparison of recognition rates between E-classifier and M-classifier

Experiment 2: classification with rejection

Results suggested in Table 3 represent ratio of rejected characters and recognition rate for accepted characters under the rejection mechanism based on posterior probability. We use 77,597 characters of 200 pages from Seungjungwon-Diary vol. 29 for testing.

Criterion (Posterior Prob.)	Ratio of Rejected Characters (%)	Recognition Rate (%)
0.9999	33.73	98.97
0.999	19.10	97.44
0.995	15.00	96.51
0.990	13.19	95.99
0.980	11.39	95.42
0.950	8.95	94.52

Table 3. The ratio of rejected characters and recognition rate

In the Table 3, under the threshold of posterior probability of 0.990, we can see the percentage of rejected characters is 13.19% and recognition rate of accepted characters is 95.99%. Also, under the threshold of posterior probability of 0.950, the percentage of rejected characters and recognition rate of accepted characters are 8.95% and 94.52%, respectively. As the threshold increases, it is reasonable that the number of rejected characters increases and recognition rate increases.

Table 4 shows economical effectiveness of proposed system with rejection mechanism. We compare the input cost by manual typing with the input cost of proposed system. Suppose that input cost and reform cost are 10 and 30, respectively, and we have 1,000,000 characters. When the ratio of rejected characters is 8.95%, total input cost of using proposed system is 25,390,000 by contrast with total cost of manual typing only method is 100,000,000.

Input Type	Total Cost	Input Cost(10 per char)
		Reform Cost (30 per char)
Manual Keying Only	100,000,000	$10,000,000 \times 10 = 100,000,000$
		0
Proposed System (Case of R.R 94.52)	25,390,000	$895,000 \times 10 = 8,950,000$
		$548,000 \times 30 = 16,440,000$

Table 4. Total cost comparison of manual keying only method and proposed method

5. Conclusions

A combination scheme between a manual typing and OCR to digitalize Korean classical materials has been proposed in this paper. In the system, a huge amount of documents was processed at once, and individual characters were identified with OCR module, and each character with the same class label was collected into a predefined group. After the grouping with OCR, an operator can verify the correctness of the classifications and finally input text codes for each group, instead of typing all the characters. With the proposed system, the typing job will not be a time-consuming labor any more and the work quality will be much better. From our preliminary examination, we believe that if the proposed system can handle more documents at once, then more characters can be inputted at the same time and its effectiveness can be dramatically increased.

References

[1] C. H. Tung, H. J. Lee and J. Y. Tsai. Multi-stage precandidate selection in handwritten Chinese character recognition system. *Pattern Recognition*, 27(8):1093–1102, 1994.

[2] C. L. Liu, Y. J. Liu and R. W. Dai. Multiresolution statistical and structural feature extraction for handwritten numeral recognition. *Fifth International Workshop on Frontiers in Handwriting Recognition(IWFHR5), Colchester, England*, pages 61–66, 1996.

[3] O. D. Trier, A. K. Jain and T. Taxt. Feature extraction methods for character recognition. *Pattern Recognition*, 29(4):641–662, 1996.

[4] S. Hara. OCR for CJK classical texts preliminary examination. *Proc. Pacific Neighborhood Consortium(PNC)Annual Meeting, Taipei, Taiwan*, pages 11–17, 2000.

[5] S. W. Lee and J. S. Park. Nonlinear shape normalization methods for the recognition of large-set handwritten characters. *Pattern Recognition*, 27(7):895–902, 1994.

[6] Y. H. Tseng, C. C. Kuo and H. J. Lee. Speeding up Chinese character recognition in an automatic document reading system. *Pattern Recognition*, 31(11):1601–1612, 1998.

[7] Y. Mizukami. A handwritten Chinese character recognition system using hierarchical displacement extraction based on directional features. *Pattern Recognition Letters*, 19(7):595–604, 1998.

[8] Y. Yamashita, K. Higuchi, Y. Yamada and Y. Haga. Classification of hand printed Kanji characters by the structured segment matching method. *Pattern Recognition Letters*, 1(8):475–479, 1983.

[9] Z. Lixin and D. Ruwei. Off-line handwritten Chinese character recognition with nonlinear pre-classification. *Proc. Inc. Conf. On Multimodal Interface(ICMI2000)*, 99(7):473–479, 2000.

Proceedings of the First International Workshop on Document Image Analysis for Libraries (DIAL'04)
0-7695-2088-X/04 $20.00 © 2004 **IEEE**