Damaged Character Pattern Recognition on Wooden Tablets Excavated from The Heijyo Palace Site

Masaki Nakagawa, Kei Saito, Akihito Kitadai, Junko Tokuno, Hajime Baba, Akihiro Watanabe

Damaged Character Pattern Recognition on Wooden Tablets Excavated from The Heijyo Palace Site

Masaki NAKAGAWA[†] *Kei SAITO*[†] *Akihito KITADAI*[†]
Junko TOKUNO[†] *Hajime BABA*[††] *Akihiro WATANABE*[††]
[†] Tokyo University of Agriculture and Technology
[††] National Research Institute for Cultural Properties
[†] nakagawa@cc.tuat.ac.jp, {kei, j-tokuno, ak}@hands.ei.tuat.ac.jp,
[††] {hajime, akihiro}@nabunken.go.jp

Abstract

This paper describes damaged character pattern recognition on wooden tablets excavated from the Heijyo palace site (the ancient palace in the Nara period from AD. 710 to 794). Since most of excavated tablets have been stained, damaged, and sometimes broken into pieces, it is extremely difficult even for archaeologists to read characters from badly blurred or missing ink on tablets. The aim of the character recognition is to output candidates even for degraded or partially lost character patterns in order to give hints to human expert readers rather than to read handwritten characters at the maximum speed. We propose a method that applies non-linear normalization and feature extraction for ternary character pattern images with gray missing area supplemented by readers. The proposed method realizes 60.2% as the 10th cumulative rate for 2,108 character patterns extracted from wooden tablets with partial damages applied artificially. This result is better than the result of the previous method.

Keywords: historical document processing, character recognition, image processing, non-linear normalization, feature extraction.

1. Introduction

A "mokkan" is a wooden tablet on which text is written with a brush in India ink. Many mokkans were used in the Nara period (from A.D. 710 to 794) in Japan. The Heijyo palace site is a ruin of the capital in the Nara period. Since the capital was moved to Kyoto after Nara, the Heijyo palace was buried under the ground and the region has been used as rice fields. Now, an enormous amount of relics have been excavated there. The soil in rice fields has been wet so that even wooden inheritances, which are fragile, oxidized or dried easily, have been kept well under the ground.

Since wooden tablets were more accessible than other media to record handwriting in the Nara period, people used them for various purposes. Up to now, 350,000 mokkans have been excavated in Japan and more than 170,000 of them are from the already excavated part of the Heijyo palace site. The number is increasing as the larger remaining part in the Heijyo palace site and other areas are excavated.

Reading handwritings on mokkans provide us the knowledge on the era. For example, decoding mokkans used as luggage tags tells the flow of materials, relations among regions, condition of economy at that period and so on.

Although we can find several preceding papers on historical document analysis [1-4], no attempt has been made to read wooden tablets excavated after more than 1,000 years.

Since most of excavated mokkans have been stained, damaged, and sometimes broken into pieces, it is extremely difficult even for archaeologists to read characters from badly blurred or partially lost ink on mokkans. Sometimes, it takes a week for archaeologists to read a single mokkan.

The goal of character recognition here is to output the correct class even for degraded or partially lost character patterns in the list of candidates so that human mokkan readers are given hints or suggestions. Therefore, *N*-cumulative recognition rate that the correct category is included within the top *N* candidates is more important than the top recognition rate.

In this paper, Section 2 describes a computer-assisted system for experts to read mokkans. Section 3 presents its image processing libraries. Section 4 describes user's augmentation of lost ink and the reason why we work on ternary images. Section 5 presents a previous character recognition method and Section 6 presents a revised method. Section 7 presents evaluation. Section 8 draws a conclusion.

2. Computer-assisted system for reading mokkans

Very often, ink on mokkans has been blurred, damaged or lost because of:

- Ink Colour has been faded out or decolored.
- Colour boundaries between ink and background (wooden tablet surface with grain) have been blurred since the surfaces of wooden tablets have been turned dark and stained.
- Mokkans are often broken and they are missing some area where characters are written.

For these reasons, experts have to make conjectures or hypotheses on the damaged or lost ink area.

We consider that image processing, handwritten character pattern recognition (HCPR) and general information processing technologies may assist the work of the experts though it is very hard to automate character recognition on mokkans.

Figure 1 shows the architecture of the computer-assisted system [5]. It consists of three components. "Image processing library" (IPL) supplies the functions of fundamental image processing. "Character recognition engine" (CRE) provides HCPR for old Japanese characters used in Nara. Mokkan-GUI or M-GUI in short is the graphical user interface and enables users to use IPL and CRE interactively. We describe each component in some detail below.

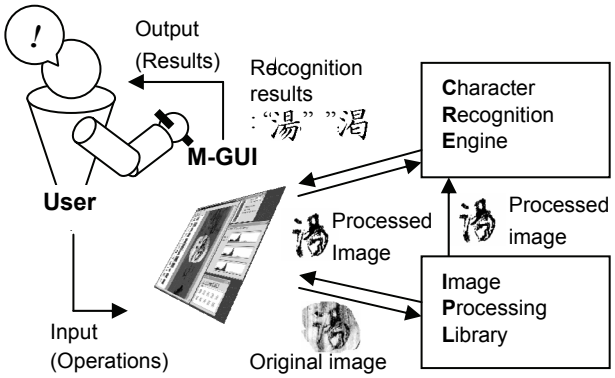

Figure 1. Architecture of the support system.

3. Image processing library

IPL has several functions to extract ink area from the digital color image of a mokkan. As the results of the extraction, IPL outputs digital binary images in which ink area is expressed as black pixels. The binary image is not only necessary for CRE but also helpful for expert users to read and decode handwriting. Also, IPL provides simple image processing methods to transform or enhance images.

Since discriminant analysis (DA) is useful for binarizing gray scale and digital color images, we employ DA as the basic algorithm of ink extraction from mokkans.

(1) DA for elemental channels of primary colors

As well as DA for gray scale images as shown in Figure 2, IPL provides DA for each elemental channel of primary colors. First, IPL generates three images by extracting each element of color channels: red (R), green (G) and blue (B). DA converts each image to a binary image. By overlapping three binary images, IPL generates an eight-colored image consisting of R, G, B, R+G, G+B, B+R, R+G+B (white), and black (black pixels in every binary image) domains. Users can remove any color domain among the eight colors to obtain the image of ink area. Figure 3 shows an example. Since the principle color element of the mokkan surface is brown, we can obtain a fine image of ink area by removing R and R+G domains.

We also employ CMY or CMYK color channels as well as RGB (C: cyan, M: magenta, Y: yellow and K: black). Figure 4 shows a successful case of ink extraction. Since mokkan surface contains yellow strongly, we can extract ink area image by adapting DA to the image of the yellow channel element.

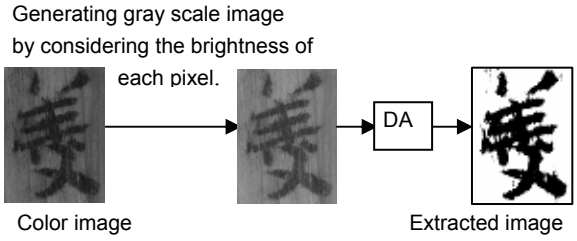

Figure 2. DA for gray scale image.

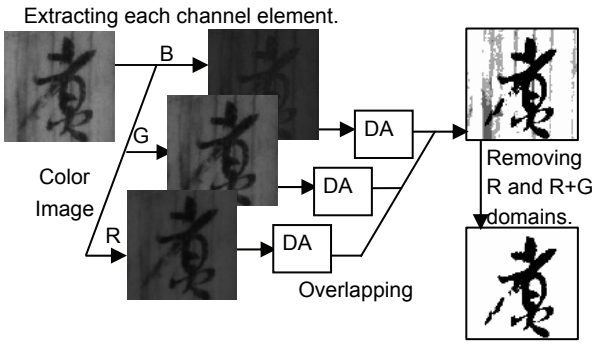

Figure 3. Ink extraction by DA for each color element.

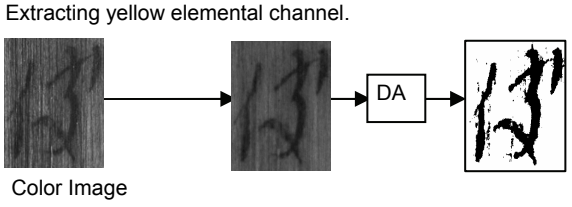

Figure 4. Ink extraction by DA for the yellow channel element in CMYK.

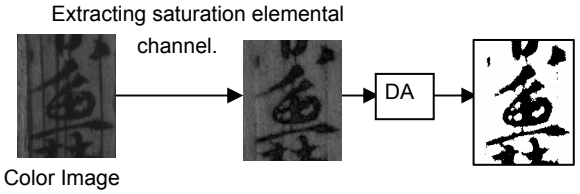

Figure 5. Ink extraction by DA for the saturation channel element in HSV.

(2) DA for elemental channels of HSV

IPL provides a function that adapts DA to each channel of hue, saturation, and value (HSV) composing the color image of a mokkan with heavy stain or dark grain. Figure 5 shows an example that DA for the elemental channel of saturation can extract ink area from

the color image of a mokkan containing extremely dark grain. Such conversions of color channels worked well for historical paper documents [1].

(3) Other image processing functions

IPL has control functions of brightness, contrast, and color balance of digital images used to generate eye-friendly images for experts.

4. Augmentation of missing ink

The character patterns on mokkans are degraded or even partially lost. The issue is to output candidates even for such character patterns. A realistic solution is for experts to augment missing ink roughly though precise augmentation cannot be expected. Therefore, we extend our HCPR method to accept a ternary image in which ink area is expressed in black, background in white and augmented area in gray.

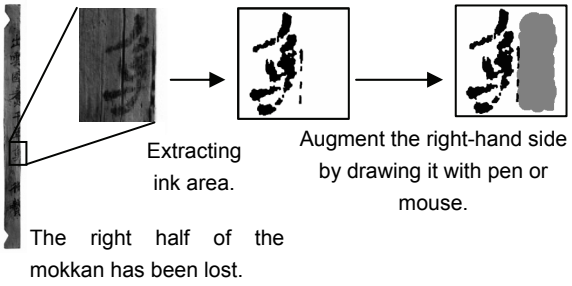

Figure 6. Augmenting information to the character pattern with missing area.

Figure 6 shows an extreme example of augmentation. Since the right half of the mokkan has been lost, all the character patterns in the mokkan may have lost their right-hand sides. In this case, the users can augment the right-hand side by drawing a rectangle in gray by pen or mouse. In other cases, the users can trace blurred ink.

Augmented gray ink has two roles. One is for non-linear normalization to normalize the original ink properly without expanding it to the whole character size as shown in Figure 7. The other role is to provide gray information to missing directional features. This is better than null features to guess original character patterns. We represent the gray level of white as 0, black as 255 and gray as 127.

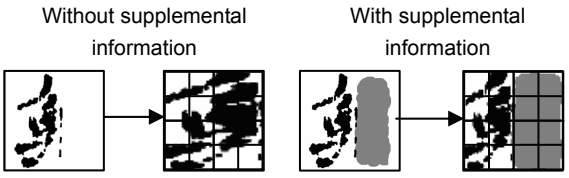

Figure 7. Nonlinear normalization with/without supplemental information.

5. Previous character recognition method

Our previous method first applies non-linear normalization to a ternary image. Second, it divides the normalized image into an array of cells. Third, it extracts 8-directional features from each cell and concatenates them into a feature vector. Fifth, it applies a discrimination function.

5.1. Non-linear normalization

First, we project pixel's gray levels of a character pattern image to X-axis and represent the distribution by $H(x)$. Similarly, we project them to Y-axis and represent the distribution by $H(y)$, respectively. Second, we compute the cumulative functions $A(x)$ and $A(y)$, respectively. Third, we non-linearly transform the image so as to linearize the cumulative functions of the transformed image as shown in Figure 8.

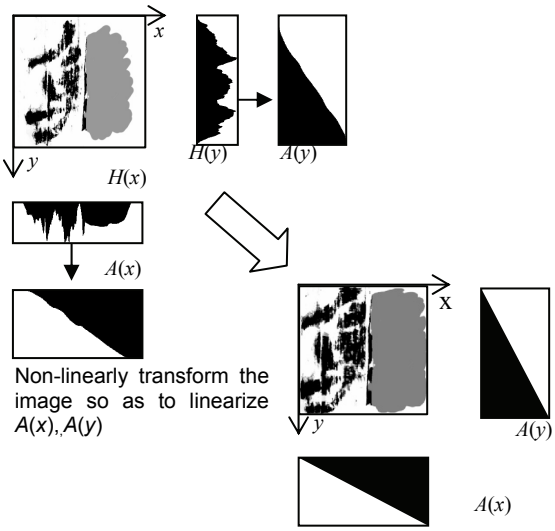

Figure 8. Previous non-linear normalization for a partially lost character pattern.

Gray levels in gray area represent the probabilities of the presence of ink so that different gray levels denote different probabilities. Therefore, dark gray area keeps its size or even expands while light gray area shrinks. Since it is very hard to guess the gray level for missing area, we change the gray level for gray area to several values and apply the above non-linear normalization with the result that multiple outputs with different sizes of the gray area are produced. For feature extraction, however, we return the changed gray levels to the standard value 127.

5.2. Feature extraction

For each pixel P_t in a non-linearly normalized ternary image, we extract 8-directional features as shown in Figure 9. The directional feature of P_t in each direction is defined by eq. (1) where d is from 0 to 7, P_d denotes one of its 8-neigbor pixels and the gray level $D(P_t)$ or $D(P_d)$ takes the value 0 (white), 127 (gray) or 255 (black).

$$f_{pixel}(P_t, d) = \left\{1 - \frac{|D(P_t) - D(P_d)|}{256}\right\} \times \{D(P_t) + D(P_d)\}$$

Then, we partition the ternary image into n x n cells and sum up each directional feature in each cell.

For each cell $C_{ab}(1<=a,b<=n)$, we make 8-dimensional feature vector $F_{cell}(a, b)$ from each directional feature $f_{cell}(a,b,d)$ as:

$$F_{cell}(a,b) = \{f_{cell}(a,b,0), f_{cell}(a,b,1), \ldots, f_{cell}(a,b,7)\} \quad (9)$$

Then, we concatenate them from all the cells into a feature vector of 8 x n x n. We apply the Gaussian filter to remove displacement distortions.

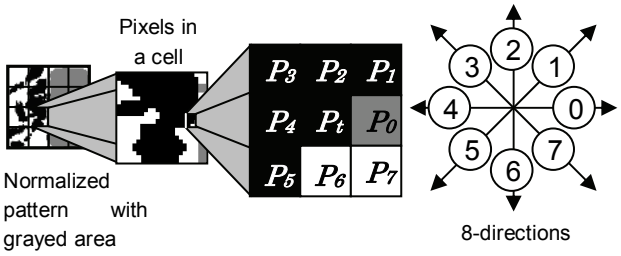

Figure 9. A pixel P_t and 8-directional features.

5.3. Discrimination function

We employ the City-Block distance between an input pattern and prototypes since the amount of learning patterns is too small to train sophisticated discrimination functions.

Since we have multiple outputs from the non-linear normalization for an input pattern, we process each output image and measure its City-Block distance to prototypes and take the minimum among the multiple outputs.

6. Revised character recognition method

In the previous method, we designed non-linear normalization and feature extraction for ternary images with gray area augmented by users. For this purpose, we employed pixel information rather than edge information being commonly used.

In the revised method, we consider to exploit the established non-linear normalization and feature extraction for binary character pattern images with as little extension as possible.

6.1. Revised non-linear normalization

We consider the case when we treat gray area as white and the case when we treat it as black and take the line density distribution for each of them. Then, we take the average of their line density distributions. Then, apply non-linear normalization so as to equalize the averaged distribution as shown in Figure 10.

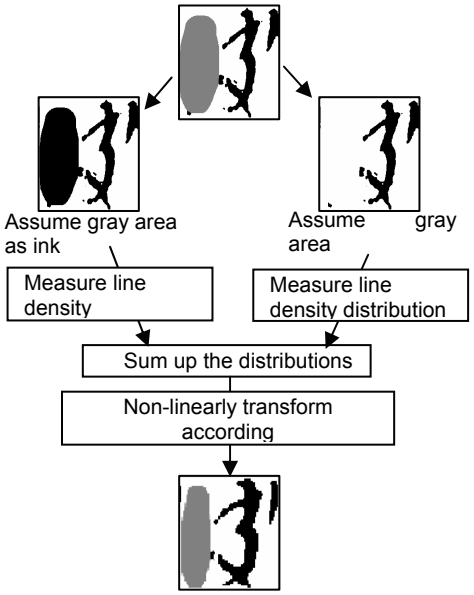

Figure 10. Revised non-linear normalization.

6.2. Non-linear normalization

Non-linear normalization, which normalizes and equalizes the distribution of stroke (line segment) density for binary character pattern images, has been proposed by several papers from the early work of Yamada [6] and Tsukumo [7]. Liu et al. combined them [8]. We extend the Liu's method so that it works on ternary images including gray area showing missing area supplemented.

First, we assume gray area as ink area and compute the stroke density functions $Hb(x)$, projection to X-axis and $Hb(y)$, projection to Y-axis. We compute the cumulative stroke density functions $Ab(x)$ and $Ab(y)$ from $Hb(x)$ and $Hb(y)$, respectively.

Second, we assume gray area as white (background) and compute the line density functions $Hw(x)$ and $Hw(y)$. Similarly to the above, we compute the cumulative stroke density functions $Aw(x)$ and $Aw(y)$ from $Hw(x)$ and $Hw(y)$, respectively.

Since the gray area may be ink or background, we take the average of the above two to obtain the cumulative line density functions $Ar(x)$ and $Ar(y)$ as follows:

$$Ar(x)=(Ab(x)+Aw(x))/2 \quad (4)$$
$$Ar(y)=(Ab(y)+Aw(y))/2 \quad (5)$$

Then, we transform the image according to the eq. (6) and (7), where the coordinate (x_{old}, y_{old}) are before the non-linear normalization and that (x_{new}, y_{new}) is after that.

$$x_{new} = x_{\max} \frac{A_r(x_{old})}{A_r(x_{\max})} \quad (6)$$

$$y_{new} = y_{\max} \frac{A_r(y_{old})}{A_r(y_{\max})} \quad (7)$$

6.3. Feature extraction

As for the feature extraction, we extract directional features for black and white areas and assume average directional features in gray area. This is similar to assuming white noise for missing packets in speech signal chopped in many time slots, which is far easier to recognize than leaving the time slots without any signal.

The directional feature extraction from character pattern contour is established. We extend this to ternary images.

Given a pixel P_t outside gray area, we compute the 4-directional features $f_{pixel}(P_t, d)$ shown in Figure 11 from pixel adjacency.

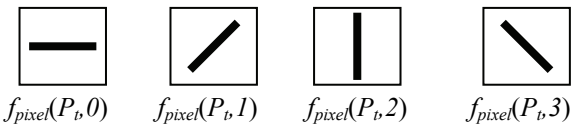

Figure 11. 4-directional features around a pixel P_t.

According to the eq.(8), we take the average of each directional feature from all the pixels except pixels in gray area and those adjacent to the gray area.

$$\overline{f_{pixel}(P_t,d)} = \frac{\sum_{P_t \in A_{bw}} f_{pixel}(P_t,d)}{k} \qquad (8)$$

Here, A_{bw} denotes the whole ink and background area and k denotes the number of pixels in A_{bw}.

For gray area and its adjacent pixels, we use these averaged 4-directional features.

Then, we partition the ternary image into n x n cells and sum up each directional feature in each cell.

For each cell $C_{ab}(1<=a,b<=n)$, we make 4-dimensional feature vector $F_{cell}(a,b)$ from each directional feature $f_{cell}(a,b,d)$ as:

$$F_{cell}(a,b) = \{f_{cell}(a,b,0), f_{cell}(a,b,1), \ldots, f_{cell}(a,b,3)\} \quad (9)$$

Then, we concatenate them from all the cells into a feature vector of 4 x n x n. We apply the Gaussian filter to remove displacement distortions.

6.4. Discrimination function

We still use the City-Block distance between an input pattern and prototypes since the amount of learning patterns is too small at this moment.

7. Evaluation

We collect sample patterns, prepare test patterns and compare the previous method and the revised method for them.

7.1. Sample patterns

We have prepared 2.108 sample binary patterns of 309 different character classes extracted from real mokkans with noise and stain removed by experts.

7.2. Preparation of test patterns

We must prepare test patterns missing some ink area. The amount of such defective patterns that are groundtruthed is small. Therefore, we use the sample patterns collected above and prepare test patterns by masking sample patterns by 8 kinds of gray masks shown in Figure 12 that represent augmented ink areas for missing areas. An example is shown in Figure 13.

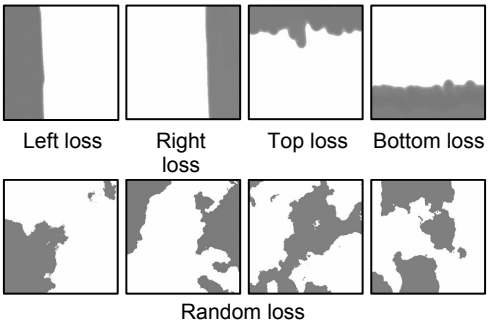

Figure 12. Masks for gray area.

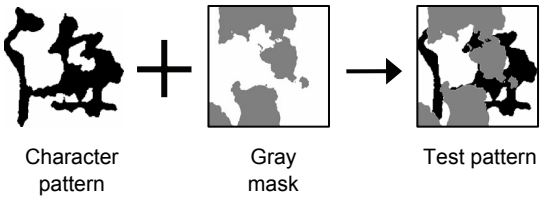

Figure 13. Preparation of test pattern.

7.3. Evaluation scheme

We employ the jack-knife method [9]. We select one pattern from the above database. We make a prototype set (dictionary) from the remaining 2,107 patterns. Moreover, we make 8 kinds of quasi partially lost patterns from the selected pattern and apply HCPR. We apply this step for 2,108 patterns, thus totally test 2,108 x 8 = 16,864 patterns.

7.4. Character recognition methods

We evaluate the following four character recognition methods made by the combination of previous or revised non-linear normalization (NN) and previous or revised feature extraction (FE).

(1) Previous NN and previous FE.
(2) Revised NN and revised FE.
(3) Previous NN and revised FE.
(4) Revised NN and previous FE.

Since the gray level of the augmented area affects the result of the previous non-linear normalization, we have prepared a 50%-decreased gray-level pattern and a 50%-increased gray-level pattern from the standard gray level 127 as well as each test pattern and applied non-linear normalization for the method (1) and (3). Then, we returned the gray level of the three output images to the standard value of 127 and applied the feature extraction.

On the other hand, the gray level of the augmented area does not affect the revised non-linear normalization, so that we have only employed the test patterns for the method (2) and (4).

The number of cells for feature extraction is set 8 x 8, so that the previous feature extraction extracts 8 x 8 x 8 = 512 features while the revised feature extracts 4 x 8 x 8 = 256 features.

7.5. Experimental results and consideration

We consider *N*-cumulative recognition rate is a good measure for our purpose. We choose 10 in this experiment. The value of *N* may be increased for difficult patterns or it can be adjusted case by case.

Table 1 shows 10-cummulative rates for the 16,864 quasi partially lost patterns. Table 2 shows them for those patterns when masked by white (background color). Table 3 shows them for 2,108 test patterns without masking.

Table 1. 10-cummulative recognition rates for quasi partially lost patterns (mask: gray).

FE / NN	previous	revised
pre.	43.2%(7282/16864)	54.8%(9249/16864)
rev.	51.2%(8641/16864)	60.2%(10145/16864)

Table 2. 10-cummulative recognition rates for quasi partially lost patterns (mask: white).

FE / NN	previous	revised
pre.	38.1%(6429/16864)	40.6%(6845/16864)
rev.	40.5%(6837/16864)	45.8%(7719/16864)

Table 3. 10-cummulative recognition rates for test patterns without masking.

FE / NN	previous	revised
pre.	66.5%(1402/2108)	70.2%(1479/2108)
rev.	70.1%(1478/2108)	74.9%(1578/2108)

The revised recognition method has achieved 60.2% 10-cummulative recognition rate, which is far better than 43.2% of the previous method. If the same feature extraction is employed, the revised non-linear normalization is superior to the previous method. Similarly, the revised feature extraction performs better than the previous one, if the same non-linear normalization is employed. Therefore, the revised non-linear normalization and feature extraction are effective for our purpose.

Moreover, both the previous method and the revised method performs better for the test patterns masked by gray color than those masked by white (background color), so that augmentation by gray seems effective for partially lost patterns as shown in table 2.

8. Conclusion

This paper presented a character recognition method for degraded or partially lost character patterns on excavated wooden tablets (mokkans). Its goal is to output correct candidates so that human expert readers are given hints or suggestions. The character recognition method applies non-linear normalization and feature extraction for ternary character pattern images with gray area showing supplemented ink by human readers. The proposed method achieved 60.2% as the 10-cumulative recognition rate for 2,108 character patterns extracted from mokkans with partial damages applied artificially. This result is better than the result of the previous method.

Acknowledgement

This work is partially supported by Grant-in-Aid for Scientific Research under the contract number S:15102001.

References

[1] Z. Shi, V. Govindaraju: Historical Document Image Enhancement Using Background Light Intensity Normalization, *Proc. 17th ICPR*, Cambridge, UK, 2aP.Mo-ii, 2004.

[2] B. Gatos, I. Pratikakis, S. J. Perantonis: An Adaptive Binarization Technique for Low Quality Historical Documents, *Proc. 6th DAS*, Florence, Italy, pp. 102-113, 2004.

[3] M-S. Kim, K. T. Cho, H.K Kwag, J. H. Kim: Segmentation of Handwritten Characters for Digitalizing Korean Historical Documents. *Proc. 6th DAS*, Florence, Italy, pp. 114-124, 2004.

[4] C. Yan, G. Leedham: Decompose-Threshold Approach to Handwriting Extraction in Degraded Historical Document Images. *Proc. 9th IWFHR*, Tokyo, Japan, pp. 239-244, 2004.

[5] A. Kitadai, K. Saito, D. Hachiya, M. Nakagawa, H. Baba and A. Watanabe: Design and Prototype of a Support System for Archaeologists to Decode Scripts on Mokkan, *Proc. 13th IGS*, Salerno, Italy, pp.54-58, 2005.

[6] H. Yamada, K. Yamamoto, T. Saito: A Nonlinear Normalization Method for Handprinted Kanji Character Recognition -Line Density Equalization-, *Proc. 9th ICPR*, pp.172-175, Roma, Italy, 1988.

[7] J. Tsukumo, H. Tanaka: Classification of Handprinted Chinese Characters Using Non-linear Normalization and Correlation Methods, *Proc. 9th ICPR*, pp.168-171, Roma, Italy, 1988.

[8] C-L. Liu, I-J. Kim, J. H. Kim: High Accuracy Handwritten Chinese Character Recognition by Improved Feature Matching Method, *Proc. 4th ICDAR*, pp.1033-1037, 1997.

[9] J.W. Tukey, "Bias and confidence in not-quite large samples," *Ann. Math. Statist.*, 29, pp.614, 1958 (abstract).