

Support System for Archeologists to Read Scripts on Mokkans

Akihito Kitadai^{*}, Kei Saito^{*}, Daisuke Hachiya^{**}, Masaki Nakagawa^{*},
Hajime Baba^{***}, and Akihiro Watanabe^{***}

^{*} Graduate School of Technology, Tokyo Univ. of Agri. & Tech. (TUAT)

^{**} Faculty of Technology, TUAT

2-24-16, Naka-cho, Koganei, 184-8588, Tokyo, JAPAN

^{***} National Research Institute for Cultural Properties, Nara
Nijyo-cho 2-9-1, Nara, 630-8577, Nara, JAPAN

Abstract

This paper describes a support system for archeologists to read "mokkan". A mokkan is a wooden tablet on which text was written by a brush. Many mokkans used in Nara period (from AD. 710 to 794) are being excavated from Heijyo-kyo, Japan (the ancient court in the Nara period). The support system is for archeologists who read mokkans that have been stained, damaged and degraded under the soil. Such mokkans are hard to read even for expert readers. However, the binarization functions of the system extract ink from the image of the mokkans and the character recognition function outputs candidates even for degraded or partially missing character patterns. We made also a graphical user interface to invoke the above functions, provide experts with suggestions and stimulate their inference. Archeologists in the experiment for evaluation enthusiastically accepted the system.

1. Introduction

In recent years, archiving cultural properties using information technologies has attracted attention.

"Mokkan" is a Japanese generic name to call a wooden tablet on which text was written by a brush in India ink. Since wooden tablets were more accessible than other media to record handwriting and they had enough weatherability, people used them for various purposes in the Nara period from A.D.710 to 794.

Heijyo-kyo is the ruins of an ancient city in Nara, Japan. It was the capital from A.D.710 to 784 and was the center of politics and economy at that time. After the capital was moved to Kyoto after Nara, Heijyo-kyo was buried under the ground and the area was used as

rice fields. Now, we can unearth many mokkans from under ground. 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 rice fields. More than 170,000 of the 320,000 mokkans unearthed in Japan come from excavations from Heijyo-kyo. The number is increasing as the excavations in the larger remaining part in Heijyo-kyo and other areas continue.

Since the first mokkan was discovered from Heijyo-kyo in 1961, the archeologists have been reading and analyzing handwritten contents on unearthed mokkans from the palace. We can acquire and extend our knowledge of this era by their work. For example, by decoding mokkans used as luggage tags, we are able to know the trade of materials, relations among regions, economy conditions at that period, etc.. The archeologists are now archiving the mokkans with their digital images and the results of the analysis.

Despite the archive of the mokkans is precious for archeology, many unearthed mokkans that have not been analyzed yet. Since most of mokkans coming from the underground have been stained, damaged and degraded, it is difficult even for experts on archeology to extract characters from badly blurred or missing ink on mokkans.

Although we find several preceding researches on information technologies for historical documents written on paper [1]-[4], no attempt has been made to use computers to process mokkans, especially such old wooden inheritances. Offline recognition methods for handwritten Japanese characters are obtaining high accuracy and robustness for blurred patterns [5], [6].

Since to read mokkans is difficult even for human experts, to seek the full automation of extracting and reading characters is impractical. However, image

processing and character recognition technologies can be used and incorporated into an interactive system which provides experts with suggestions and stimulate their inference.

This paper describes the support system that helps archeologists to read mokkans. Section 2 presents the support system with the information processing to read mokkans. Section 3 shows the opinions by archeologists in an experiment for the evaluation of the support system. Section 4 draws conclusion.

2. Support system using information processing technologies

2.1. Basic idea of a support system

To read handwriting on a mokkan, experts extract ink parts from the mokkan or its picture first. However, very often, ink has been blurred, damaged or missing because:

- Color of ink parts has been faded out or decolorized.
- Color boundaries between ink parts and the background (skin of wood with grain) have been vague since the surfaces of wooden tablets have become darkish and stained.
- Some regions of a mokkan containing ink parts have been lost or broken.

For these reasons, experts have to make conjectures or hypotheses on the missing ink parts.

We consider that image processing, handwritten character pattern recognition (HCPR) and some technologies of information processing may assist the experts' work. At least, the results of the information processing can stimulate the inference of the experts.



Figure 1. Mokkans unearthed from Haijyo-kyo

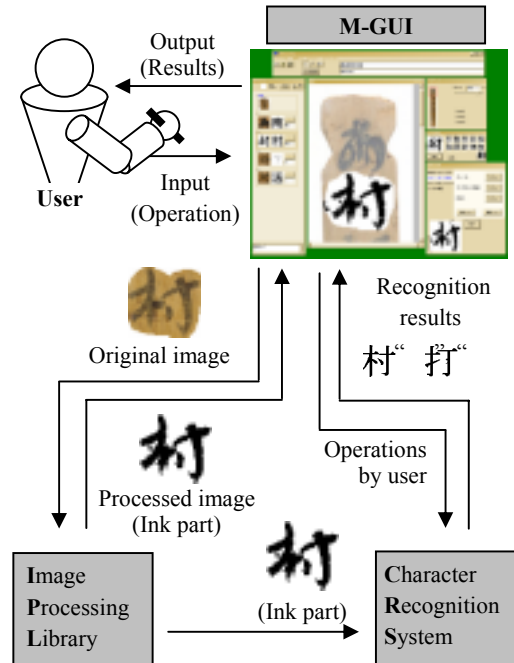


Figure 2. Architecture of the support system

2.2. Architecture of the support system

We show the architecture of the support system in Figure 2. The support system 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 it enables users to use IPL and CRE interactively. We describe the details of each component below.

2.3. Components of the support system

IPL has several functions to extract ink parts from the digital color image of a mokkan (Figure 3, 4, 5 and 6). As the results of the extraction, IPL outputs digital binary images in which ink parts are expressed as black pixels. We have used discriminant analysis (DA) as the basic algorithm of ink extraction from mokkans with darkish grain, stained surface, and ink parts faded out. The binary image is not only necessary for CRE but also helpful for expert users to read and decode handwriting on a mokkan. Also IPL supplies control functions of brightness, contrast, and color balance of digital images used to generate eye-friendly images for experts.

CRE in this system provides character recognition process for 241 categories commonly used in the mokkans. Since the total amount of the categories in the mokkans are about 1,400, we are expanding the dictionary of the templates. CRE should output candidates even for degraded or partially missing character patterns and stimulate imagination of experts. The most important problem is to output candidates even for character patterns with missing ink parts. A realistic solution is that experts roughly mark the missing ink parts of the character (we call this process

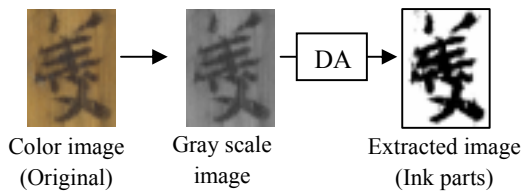


Figure 3. DA for gray scale image

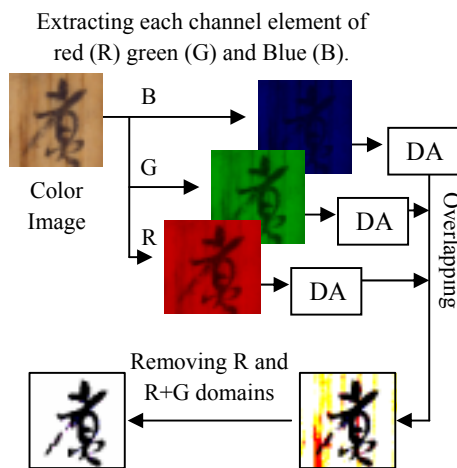


Figure 4. Ink parts extraction by using DA for each color element

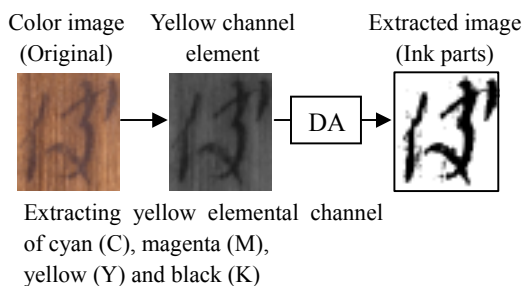


Figure 5. Ink extraction by using DA for the yellow channel element in CMYK

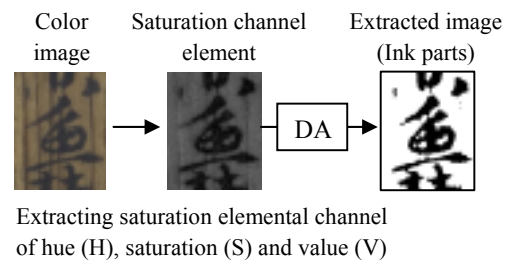


Figure 6. Ink parts extraction by using DA for saturation channel element in HSV

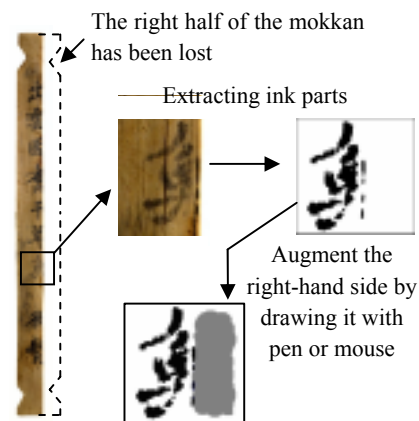


Figure 7. Ink extraction by using DA for the yellow channel element in CMYK

augmentation). Therefore, we extended our HCPR method to accept a gray scale image in which ink parts are expressed in black, background in white and augmented parts in gray.

Figure 7 shows an extreme example of argumentation. Since the right half of the mokkan has been lost, we can conjecture that all character patterns in the mokkan have lost their right-hand sides. In this case, the users can only augment the right-hand side by drawing a rectangle in gray by pen or mouse. In other cases, the users can restore the blurred or damaged ink by tracing supposed strokes.

Our CRE first apply nonlinear normalization to a gray scale image. Second, it divides the normalized image into an array of cells. Third, it extracts 8-directional features from each pixel as shown in Figure 8 [7]. For each pixel P_i , the directional feature of P_i in each direction $F(P_i, d)$ is defined by eq. (1) where d is from 0 to 7, P_x denotes P_i or one of its 8-neighbor pixels P_d and $C(P_x)$ represents the color density of P_x . $C(P_x)$ takes the value 0 (white), 255 (black), or between 1-254 (gray). Fourth, within each cell, the

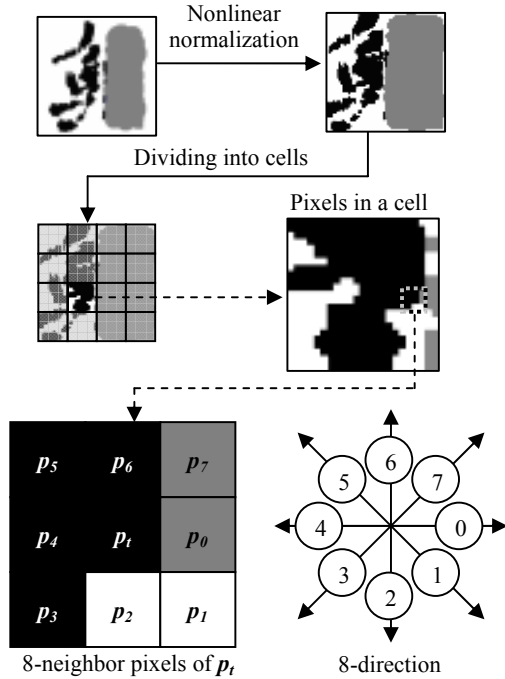


Figure 8. Pixels in a normalized pattern and 8-directions

directional feature $F(P_i, d)$ of each pixel P_i is summed up for every direction.

$$F(P_i, d) = \left\{ 1 - \frac{|C(P_i) - C(P_d)|}{255} \right\} \times \{C(P_i) + C(P_d)\} \quad (1)$$

Augmented gray ink has two roles. One is for nonlinear normalization to normalize the original ink parts properly without expanding it to the whole character size as shown in Figure 9. As a result, each directional feature will be extracted from the cell at correct place. The other role is to provide white noise to missing directional features. This is better than null features to guess the original character patterns. Table 1 shows examples in which gray ink works well.

Since CRE performs the nonlinear normalization and feature extraction regarding the color density of gray ink, the users can hypothesize larger amounts of missing ink by using a darker gray at the same time that lighter gray represents smaller amounts of them (Figure 9). Also, before nonlinear normalization, CRE can create multiple character patterns with different density of gray ink from an input pattern automatically in order to output candidates for various hypotheses of the amount of missing ink.

In contrast with the human process of character recognition that regards the characteristic features in a

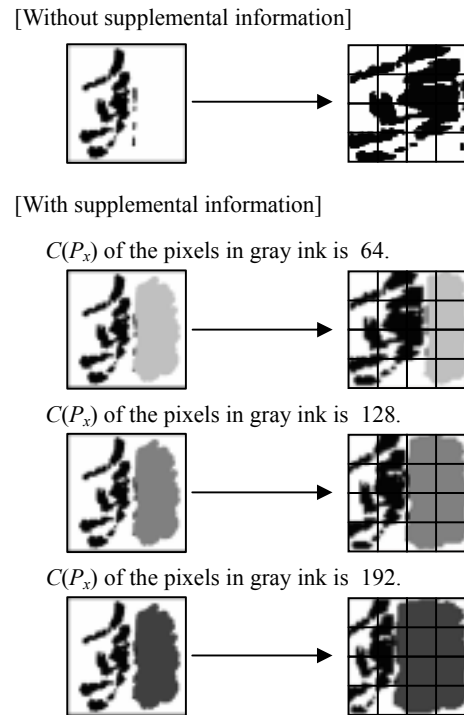


Figure 9. Nonlinear normalization with/without supplementary information

pattern, machine process evenly use whole features extracted from character patterns. Since the difference between the processes, we can expect that CRE outputs candidates helpful for expert readers of mukkans.

M-GUI shown in Figure 2 provides the functions of IPL and CRE to users of the support system. The user of the support system can perform most of operations by using pen or mouse. Moreover, it manages the process of reading and decoding handwriting on mukkans. Since M-GUI can save, manage, and load the workspaces consisting of HCPR results, edited images, etc., the users of the support system can suspend and resume their work at any time and can share the workspaces for collaborative work. Also, M-GUI can export the HCPR results and edited images to word processors, image editors and other application software.

3. Evaluation by archeologists

To evaluate the support system, we performed an experiment in which two archeologists participated. By the experiment, we show some opinions by them:

- The ink on mukkans becomes visible clearly by the

binarizing functions provided by the system.

- Choosing the best binarizing functions manually is difficult.
- Since the system outputs different candidates of character recognition by supplementing and changing gray ink, we can perform the analysis of the mukkans considering various hypotheses of the missing ink.
- We need more categories of character supported by the recognition process of the system.

Since the system can export the edited images and candidates of character recognition to other software e.g. word processor, we can use the system not only for the process of analyzing and reading mukkans but also for writing reports and documents about analyzed mukkans.

4. Conclusion

We described a support system that helps experts to read mukkans. Ink extraction by image processing and candidate suggestion by character recognition are the main results. Since we are now constructing a pattern database of old characters, the expansion of the categories of character and the evaluation of the character recognition of the system are our future work. Our future research also include to develop an

automatic method for choosing the binarizing functions and to consider recognition methods in which so-called “gradient feature” or “gradient direction” is used efficiently [6], [8].





5. Acknowledgment

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

6. References

- [1] B. Gatos, I. Pratikakis, and S.J. Perantonis, “An Adaptive Binarization Technique for Low Quality Historical Documents”, *Proc. 6th Workshop on Document Analysis Systems*, Florence, Italy, Sep. 2004, pp. 102-113.
- [2] M.S. Kim, K.T. Cho, H.K. Kwag and J.H. Kim, “Segmentation of Handwritten Characters for Digitalizing Korean Historical Documents”, *Proc. 6th Workshop on Document Analysis Systems*, Florence, Italy, Sep. 2004, pp. 114-124.
- [3] C. Yan and G. Leedham, “Decompose-Threshold Approach to Handwriting Extraction in Degraded Historical Document Images”, *Proc. 9th IWFHR*, Tokyo, Japan, Oct. 2004, pp. 239-244.
- [4] Z. Shi and V. Govindaraju, “Historical Document Image Enhancement Using Background Light Intensity Normalization”, *Proc. 17th ICPD*, Cambridge, UK, Aug. 2004, 2aP. Mo-ii.
- [5] M. Suzuki, N. Kato, H. Aso and Y. Nemoto, “A handprinted character recognition system using image transformation based on partial inclination detection”, *IEICE Trans. Inf. & Syst.*, May. 1996, vol. E79-D, no. 5, pp. 504-509.
- [6] K. Sawa, T. Wakabayashi, S. Tsuruoka, F. Kimura and Y. Miyake, “Accuracy Improvement by Gradient Feature and Variance Absorbing Covariance Matrix in Handwritten Chinese Character Recognition”, *IEICE Trans.*, Nov. 2001, vol. J84-D-II, no. 11, pp. 2387-2397 (in Japanese).
- [7] C.L. Liu, Y.J. Liu and R.W. Dai, “Preprocessing and Statistical/Structural Feature Extraction for Handwritten Numeral Recognition”, *Progress of Handwriting Recognition*, A.C. Downton and S. Impedovo eds. World Scientific, 1997, pp. 161-168.
- [8] C-L. Liu, K. Nakashima, H. Sako, H. Fujisawa, “Handwritten digit recognition: benchmarking of state-of-the-art techniques”, *Pattern Recognition*, Oct. 2003, vol. 36, no. 10, pp. 2271-2285.

Table 1. Recognition results with/without supplementary information

Character pattern	Recognition result
With supplemental information 	The correct code “良” is as the 7th candidate
Without supplemental information 	The correct code “良” is not in the 1-10th candidates
With supplemental information 	The correct code “麻” is as the 4th candidate
Without supplemental information 	The correct code “麻” is not in the 1-10th candidates