Proc. 8th International Workshop on Document Analysis System (DAS 2008),
Nara, Japan Vol I, pp.533-538 (2008.9).

Document Image Retrieval to Support Reading Mokkans

Akihito KITADAI, Jun TAKAKURA, Masatoshi ISHIKAWA, Masaki NAKAGAWA
Tokyo University of Agri. & Tech.
a.kitadai@gmail.com

Hajime BABA, Akihiro WATANABE
National Research Institute for Cultural Properties, Nara.
hajime@nabunken.go.jp

Abstract

This paper presents a design and an implementation of document image retrieval to support reading mokkans. A mokkan is a wooden tablet with text written by a brush in India ink. Despite the archaeological and historical value of the mokkans excavated from ancient ruins, many of the mokkans have not been decoded yet due to the lost or too much damaged character patterns on them. Character recognition for damaged patterns is useful to decode such mokkans. Furthermore, if the recognition results show not only the character codes but also the images of the character patterns and the whole mokkans, the recognition becomes useful document retrieval to complement the lost or unreadable part of the mokkans. In the implementation, we built a public database of historical mokkans with their photographs and a character recognition module working on our support system to search the database. The evaluation by archaeologists is in progress.

1. Introduction

"*Mokkan*" is a Japanese generic name to call a wooden tablet with text written by a brush in India ink. There are about 320,000 old *mokkans* excavated at ruins of ancient cities in Japan. Figure 1 shows some of the *mokkans* from the ruin of the *Heijo* palace site, the capital of Japan in the Nara period (from A.D. 710 to 794). Many *mokkans* were used for luggage tags of gifts, commodities, goods for tax, and so on. Therefore, decoding these *mokkans* is important to find the flow of materials, the relations among regions and the condition of economy at the period.

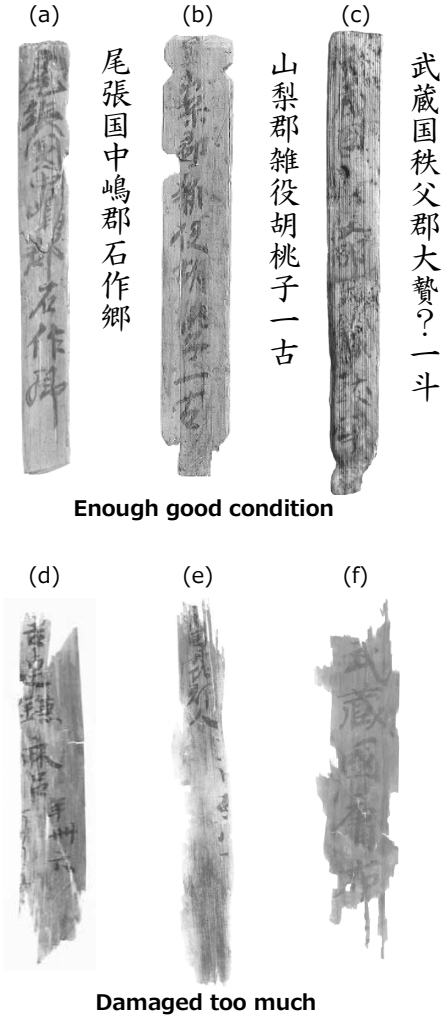

Figure 1. Mokkans excavated from Heijyo palace site.

In spite of the archaeological value of the *mokkans*, only a little amount of them have been decoded and most of them are remaining to be analyzed. Since most of the excavated *mokkans* from the ruins have been stained and damaged (e.g. (d)-(f) in figure 1), it is difficult even for archaeologists to extract characters from badly blurred or missing ink on *mokkans* and read them. The number of undecoded *mokkans* is increasing owing to excavations while the number of experts is decreasing.

Recently, some research proposed supporting methods and systems to decode historical documents [1]-[3]. We also proposed a system to help archeologists decode *mokkans*. This system employs image processing, character recognition and information retrieval [4].

2. Details and problems in decoding *mokkans*

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

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

For these reasons, the archeologists have to make conjectures or hypotheses on the lacked ink parts.

The archeologists had been using commercial image processing software for decoding when needed. For example, they have been used “Adobe Photoshop”. However, this makes decoding work interleaved and disturbed. Powerful image processing suitable for *mokkans* must be incorporated into the support system.

Another problem with decoding *mokkans* is that the archaeologists have to consult many dictionaries and books on handwritten character patterns, place names, people names and so on as shown in Figure 2. They are also interruptive for them. The support system should help them consult appropriate information without interrupting their thinking.

Yet another problem is that the archeologists draw pictures to decode a mokkan as shown Figure 3. Drawing and writing is part of decoding so that the support system should provide pen interfaces.

Figure 2. Mess of documents necessary for decoding.

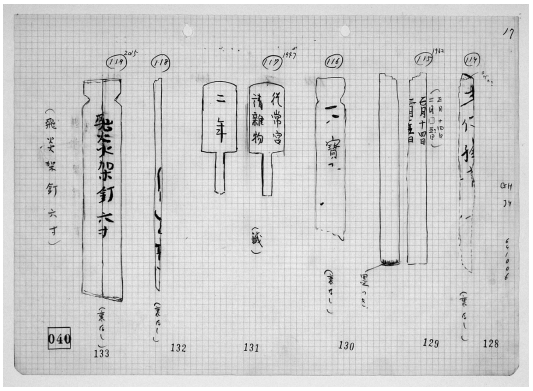

Figure 3. Drawing for decoding *mokkans*.

3. Traditional technologies

3.1. Support system to decode *mokkans*

To support the archaeologists, we have proposed a support system for decoding *mokkans* (Figure 4) [4]. The system provides the functions of image processing, character recognition, context processing and communication between archaeologists. The image processing functions are useful to extract ink on degraded and darkish surfaces of the *mokkans* (Figure 5). The character recognition function recognizes character images even though some parts of the ink are missing (Figure 6) [5]. The context processing function helps to complement decoding results of damaged *mokkans*. And the communication function supports the collaborative work among the archaeologists because two heads are better than one in decoding *mokkans* (Figure 4).

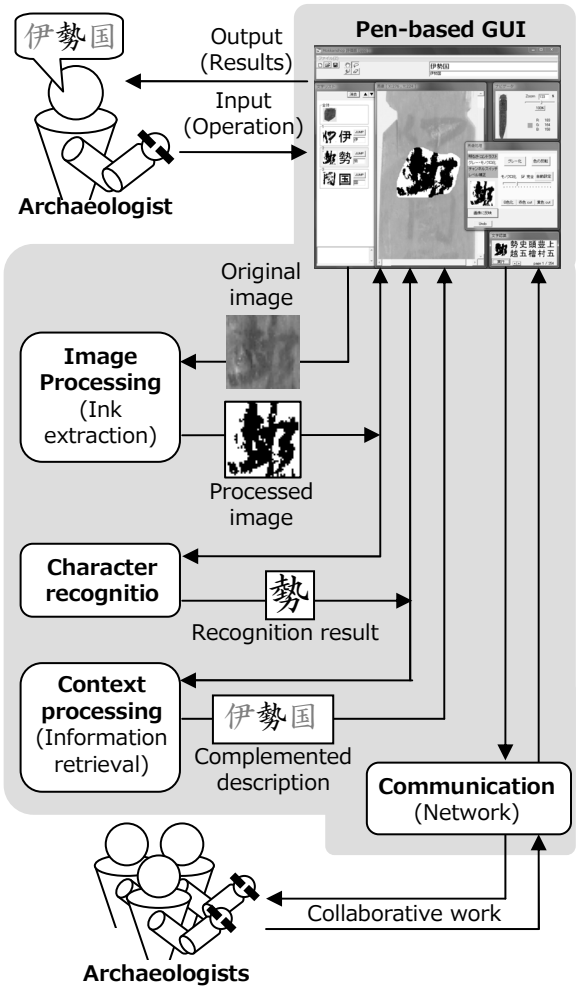

Figure 4. Support system to decode *mokkans*.

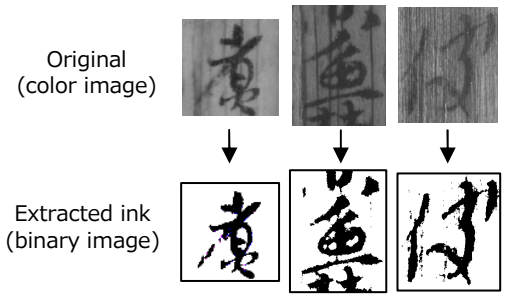

Figure 5. Examples of image processing.

The archaeologists can chose the pen-based graphical user interface of the support system between "multi-window type" and "tiling-window type" (Figure 7). Both of them invoke the above functions, provide the archaeologists with suggestions and stimulate their inference [6].

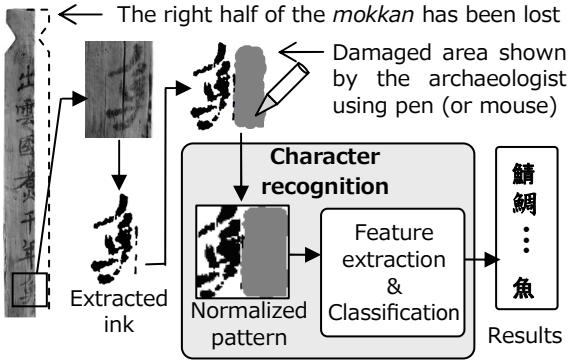

Figure 6. Process of character recognition for damaged character pattern.

The screen shot of the system with multi-window interface

The screen shot of the system with tiling-window interface

Figure 7. Two types graphical user interface of support system.

The support system has been already supplied to the archaeologists to support their work.

3.2. Database for archaeologists

We can find similar descriptions among multiple *mokkans* because the number of *mokkans'* usage was limited and the length of the description was less than a hundred characters. This fact shows that a *mokkan* might be useful to decode other *mokkans* and large databases of *mokkans* are helpful for the archaeologists.

For that reason, the archaeologists try to search *mokkans* that have similarities with the damaged *mokkan*. Real *mokkans*, analog/digital photographs and books carrying *mokkans* are referred by the archaeologists. However, the number of the references is too large. The archaeologists need methods to search the references easily and quickly.

To respond to the request, we have opened a database of the *mokkans* to public via our website [7]. The database indexes the *mokkans* that have been decoded partially or completely. The archaeologists can obtain the information of the *mokkans* (decoded description, category of shape, place of discovered/stored and so on) (Figure 8).

Figure 8. Public database of *mokkans* for archaeologists.

By constructing and opening the database, the work of reference becomes easier and quicker than the past. However, the database provides only "text-based" indexes. It shows the fact that we have no method to search the database by images of the *mokkans*. This is a critical problem of the database because text information extracted from damaged *mokkans* does not have enough reliability.

4. Design

4.1. Basic idea

We consider that a starting point to solve the above problem is to provide document image retrieval that accepts a damaged character image as a search key and returns whole *mokkan* images containing similar character images to the search key. This is the motivation behind this research.

Figure 9 shows the basic idea of the document image retrieval to support reading *mokkans*. The search key of the retrieval is a damaged character images on a decoding target *mokkan*. After the search engine that is built in our support system returns character images that are similar to the key, the archaeologists can choose one of them to refer the whole *mokkan* image that owns the character image.

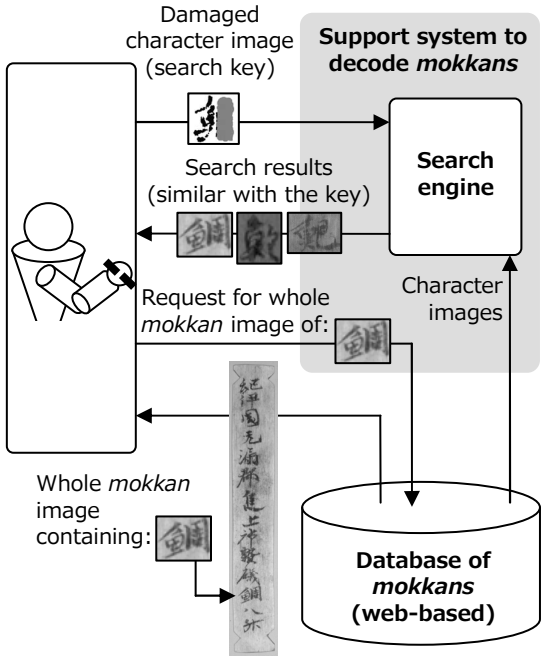

Figure 9. Basic idea of document image retrieval.

4.2. Improved database

We redesigned the database for the document image retrieval method. The new database provides both the whole images and discrete character images of *mokkans*. The former images are good for referring the writing style, context and shape of the *mokkans*. The latter images are not only good for referring character shapes but also necessary for providing the similar character images to the search key.

Also, the new database provides links between the character images and the whole *mokkan* images to show which *mokkan* owns the character image, positional information on the owner *mokkan* to show the written place on the owner *mokkan*, and the pre-extracted feature vector from each character image accelerates the speed of search engine. We collectively call the link, positional information and feature vector as "profile" of the character image.

4.3. Search engine

We can use a character recognition function as the search engine for document image retrieval when it can evaluate the similarity between character images on the documents. In this research, we employed the character recognition function for damaged character images (presented in Section 3.1). The 10th accumulative rate of the function for 16,864 quasi damaged character images is about 60% [5].

5. Implementation

5.1. Profile editor

We built the "profile editor" shown in Figure 10. The main roles of it are extracting feature vectors from character images and appending profiles (including the feature vectors) to the character images.

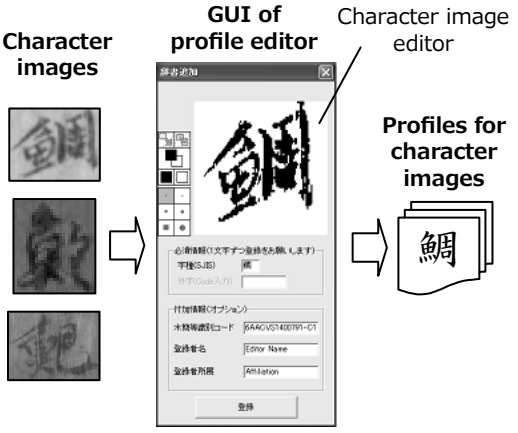

Figure 10. Profile editor.

The profile editor works together with our support system above mentioned. Therefore, by using the image processing functions of the support system, the archaeologists can clip character images from whole *mokkan* images and remove their noise due to the stains and wood grain on the *mokkans* before extracting feature vectors. Also, the profile editor automatically inserts the links between character images and whole *mokkan* images when the archaeologists clip character images on the support system.

Additionally, when damaged parts of a character image are recoverable, the archaeologists can supplement the missing ink on the profile editor for high accuracy retrieval.

5.2. Public database

Figure 11 shows the web interface of the new database that provides whole and discrete character images of the *mokkans* [8].

Text-based document retrieval is available on the web site. However, we provide the document image retrieval method via the support system for the quick response and reducing the burden of the web server.

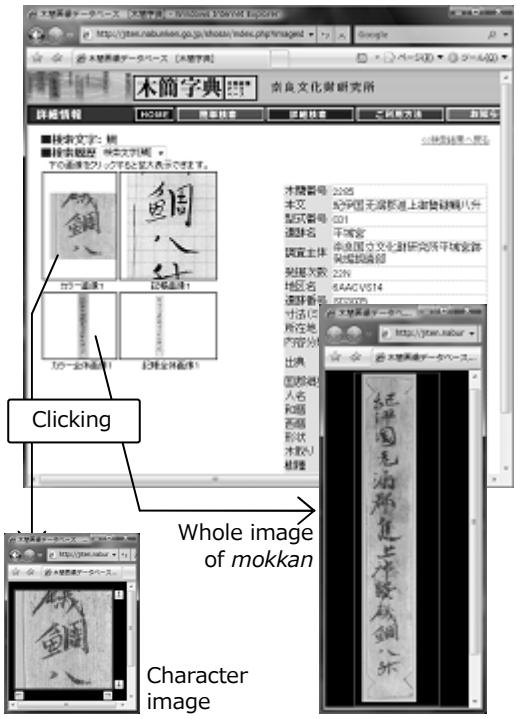

Figure 11. Database providing images.

5.3. Search engine by character recognition function

The search engine built in the support system accepts a damaged character image as the search key, and provides the list of its similar character images evaluated by the character recognition method. By

selecting one of the similar character images, its whole *mokkan* image is displayed on the web interface of the support system (Figure 12).

For the visibility of the archaeologists, the graphical user interface of the search engine displays the character images as binary. Its color image is provided by another web interface of the support system.

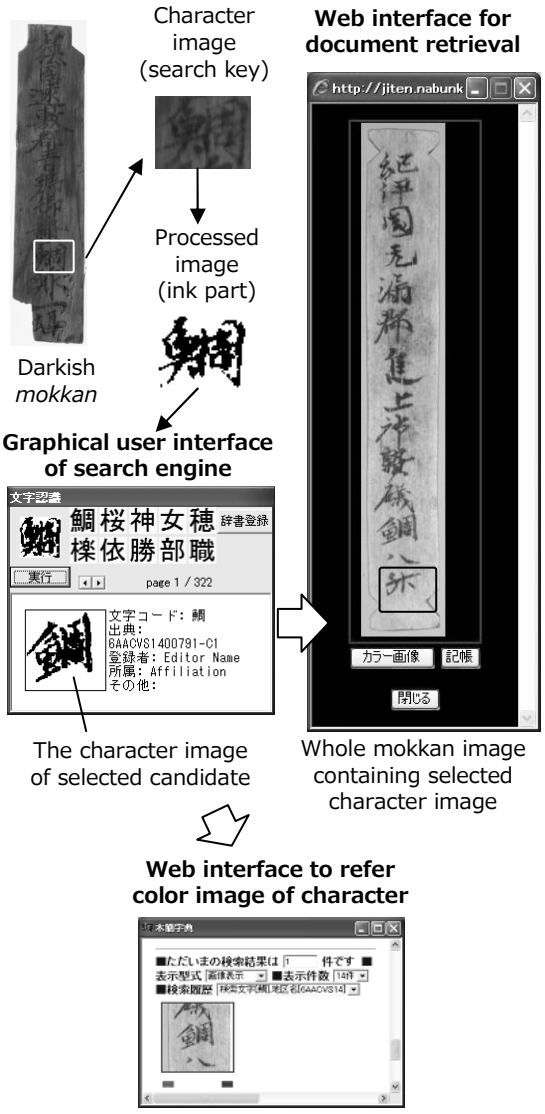

Figure 12. Search engin to provide whole mokkan image.

6. Conclusion

In this paper, we presented the design and implementation of document image retrieval to support reading *mokkans*. The evaluation of the performance and effectiveness by the archaeologists are in progress.

The research of the method that accepts a sequence of character images as a search key is one of our future work. The adaptation of our method to other historical documents is interesting for us.

7. Acknowledgement

This work is being supported by the grant-in-aid for scientific research under the contract number S-20020001 and the grant-in-aid for young scientists under the contract number B-19720202.

8. References

[1] J. He, Q. Do, A. Downton and J. Kim, "A Comparison of Binarization Methods for Historical Archive Documents", *8th International Conference on Document Analysis and Recognition*, Seoul, Korea, Aug. 2005, pp. 538-542.

[2] B. Gatos, I. Pratikakis and S.J. Perantonis, "An Adaptive Binarization Technique for Low Quality Historical Documents", *Proc. 6th International Workshop on Document Analysis Systems*, Florence, Italy, Sept. 2004, pp. 102-113.

[3] M.S. Kim, K.T. Cho, H.K. Kwag and J.H. Kim, "Segmentation of Handwritten Characters for Digitalizing Korean Historical Documents", *Proc. 6th International Workshop on Document Analysis Systems*, Florence, Italy, Sept. 2004, pp. 114-124.

[4] A. Kitadai, Y. Tone, M. Ishikawa, M. Nakagawa, H. Baba and A. Watanabe, "Support System for Standalone and Collaborative Work of Archaeologists to Decode Mokkans", *Proc. 13th Conference of the International Graphonomics Society*, Melbourne, Victoria, Australia, Nov. 2007, pp. 226-229.

[5] M. Nakagawa, K. Saito, A. Kitadai, J. Tokuno, H. Baba and A. Watanabe, "Damaged Character Pattern Recognition on Wooden Tablets Excavated from The Heijyo Palace Site", *Proc. 10th International Workshop on Frontiers in Handwriting Recognition*, La Baule, France, Oct. 2006, pp. 533-538.

[6] Y. Tone, A. Kitadai, M. Ishikawa, M. Nakagawa, H. Baba and A. Watanabe, "User Interface Design for a Mokkan Reading Support System", *Proc. 13th Conference of the International Graphonomics Society*, Melbourne, Victoria, Australia, Nov. 2007, pp. 193-196.

[7] http://www.nabunken.jp/Open/mokkan/mokkan2.html

[8] http://jiten.nabunken.go.jp/