A System to Annotate and Cluster Pieces of Mokkan

Kha Cong Nguyen, Truyen Van Phan, Masaki Nakagawa
Department of Computer Science
Tokyo University of Agriculture and Technology
Email:congkhanguyen@gmail.com, truyenvanphan@gmail.com, nakagawa@cc.tuat.ac.jp

***Abstract*—Mokkan is a generic name of wooden tablets on which characters are written with brush and ink. They were used as gift tags, message boards, documents and so on in ancient periods in Japan. Over the long-term history with extreme impacts of climate and environment, however, most of unearthed pieces of mokkan have been stained, damaged, degraded and broken. As a result it is extremely difficult even for archaeologists to read characters from badly blurred or missing ink on mokkan. In this paper, we describe a system to annotate information on mokkan pieces excavated from Heijyo-kyo palace, Nara, Japan and to cluster them based on the annotated information, shape, size and color features. Here, we apply the most advanced technologies of handwriting recognition, image processing methods to support users to annotate.**

***Index Terms*—Mokkan, image processing, handwriting recognition, annotation, clustering.**

I. Introduction

The value of history and culture is invaluable for every country. Mokkan is a generic name of wooden tablets used as documents in ancient time in Japan. Until now, the total number of mokkan excavated in Japanese ruins is about 320,000 pieces and 170,000 pieces or more of them are excavated from the ruins of the Heijyo-kyo palace site in Nara [1]. Although the Nara period is approximately one century, along with the strong influence of China and Buddhism, this period has an important meaning in the Japanese history. Therefore, decoding mokkan is necessary to understand this period deeply and clearly. From mokkan, we could learn flow of materials, relations among regions, condition of economy and so on.

Fig. 1: The plate of mokkan from the Heijyo-kyo palace site.

Nevertheless, most of excavated pieces of mokkan are damaged, stained and sometimes broken into pieces so that character images are rarely kept completely. Another difficulty with decoding mokkan is that archaeologists have to consult many dictionaries and books on character patterns, spend much time finding their previously related comments or their colleagues comments on the character patterns.

Due to the development of Information Technology (IT), however, some of the difficulties can be overcome and new functions can be provided. First, mokkan images can be stored digitally and shared by many researchers or even ordinary people. Image processing methods can be applied to restore damaged images [3], handwriting recognition technology can be used to suggest candidate classes for character patterns difficult to read [5], and databases can be used to store and access mokkan images to support archaeologists.

Previously, we made a system named MokkAnnotator to achieve mokkan images on a glass plate having about 10 pieces as shown in Figure 1 and apply IT as mentioned above [2]. Although it has been effectively used in the Nara National Research Institute for Cultural Properties (NNRICP), however, MokkAnnotator did not allow users to manage each piece of mokkan and classify mokkan pieces based on the content of annotations. Therefore, we propose a new system named MokkAnnotator II for archaeologists to manage each mokkan piece, add annotations, share information between users and classify mokkan pieces based on the content of annotations.

The remaining of the paper is organized as follows: section II shows the system architecture, section III presents the GUI and section IV draws conclusion.

II. The overview of MOKKANNOTATOR II

This section describes the input, output and main functions of the system as shown in Figure 2. The functions are designed through the discussion with the archaeologists at NNRICP.

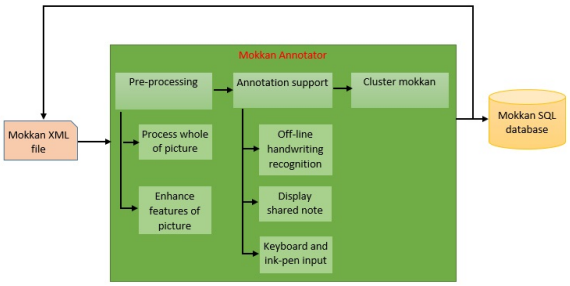

Fig. 2: The proposed functions of the MokkAnnotator II

A. Input and output

In order to manage the information of each mokkan piece, users, annotation layouts (mokkan image) and annotations, we create an input data file for each mokkan piece in XML (Extensible Markup Language) as shown in Figure 3, including following nodes:

```
<?xml version="1.0"?>
<Mokkan mokkan_ID='1' grid_number='10008' bat_number='1228' excavated_time='2010/10/1' glass_number='123'>
  <Users>
    <User user_ID='1' user_name='CongKhaNguyen' department='TUAT' age='25' sex='Male' />
    <User user_ID='2' user_name='Wantanabe' department='NNRICP' age='55' sex='Male' />
  <Images>
    <Image image_ID='1' type='colorRGB24' path='..\mokkan1_1.jpg' created_time_date='2010/10/2' user_ID='1'>
      <Annotations>
        <Annotation annotation_ID='1' created_time_date='2015/1/1' user_ID='2' type='decoding'>
          <position first_point='400,120' width='175' height='144' shape='rectangle' />
          <Content>"国"かもしりません </Content>
          <Ink_coordinate> 12.25507853377791,36.07954952145647... </Ink_coordinate>
        <Annotation annotation_ID='2' created_time_date='2015/1/1' user_ID='1' type='shape_feature'>
          <Position first_point='55,200' width='66' height='66' shape='star' />
          <Content>081型式 </Content>
    <Image image_ID='2' type='Binary' path='..\mokkan1_2.bmp' created_time_date='2010/10/2' user_ID='2'>
      <Annotations>
        <Annotation annotation_ID='1' created_time_date='2014/10/1' user_ID='2' type='other'>
          <Position first_point='70,10' width='70' height='25' shape='rectangle' />
          <Content>同じ里からの木簡アリ『平城宮』1-1</Content>
        <Annotation annotation_ID='2' created_time_date='2014/10/15' user_ID='1' type='shape feature'>
          <Position first_point='25,400' width='70' height='40' shape='circle' />
          <Content>未調整？ </Content>
```

Fig. 3: Mokkan piece XML file

1) Mokkan node: in NNRICP, unearthed mokkan pieces are stored in trays called bats. Each bat includes several glasses to place mokkan pieces excavated in the same area called grid in Heijyo-kyo palace. Mokkan information is mokkan id, grid number, bat number, excavated time, and glass number. With the information, it is easy to know where a piece of mokkan is stored in the mokkan repository and in which grid in Heijyo-kyo palace the piece has been excavated.
2) User nodes: include the information of the user such as user ID, user name, department, age, and sex. Each time, a user logs in to the system, a new session is established. When a user creates a new mokkan image and adds a new annotation, user ID will be saved to the correspondent image and annotation node in the XML input file. Users can modify the annotation by others, so that user ID is also included in annotation nodes to show who the annotation belongs to.
3) Image nodes (annotation layout node): include image ID, its type(color, gray scale, binary), the path of the image, user ID who has created it, modified it, with created time and date and the image processing methods used (described more clearly in the sub-section II-B).
4) Annotation nodes: contain annotation ID, user ID, created time and date, type, position of comment text-boxes (width, height, the first point and shape) and handwriting coordinates. The above attributes will be saved to the input XML file for each mokkan piece. The content of each annotation will be filtered and saved to SQL database as a mokkan dictionary. Generally, the mokkan dictionary will contain some fields such as ID, mokkan ID, user ID, the type of annotation, content and category (indicating the group it belongs to after clustering). With decoding comments, radical, number of strokes, Unicode, and meaning explanation fields are also included. In some cases, the content of annotations is unstructured, it cannot be filtered and put all to the content field. Users can find annotations by user ID, mokkan ID, radical, Unicode and so on.

After the system is closed (the annotation process is completed), the attributes of newly created and modified annotations, annotation layouts and user information will be updated into the input XML file and mokkan SQL database.

B. Mokkan image pre-processing

As we mentioned above, since mokkan have been buried under-ground for more than 1300 years, the written ink has become worn out, it is difficult for archaeologists to decode characters on them. Therefore, the system provides three types of image: color, gray scale and binary images as well as the two following methods to adjust the properties of images [3]. Each type of image and method will highlight different details of the image:

1) Methods to enhance the readability of whole mokkan image such as noise reduction using the Gaussian filter, and adjusting the contrast and brightness of image.
2) Methods to enhance some selected pixels or border of image.

After users complete the image adjustment, a layout of the adjusted image will be created and users can freely annotate on this layout. The information of image such as type, used image processing methods, and user ID are also updated to XML file.

C. Annotation support

There are three types of comments as shown in Figure 4 which users can add to a mokkan annotation layout:

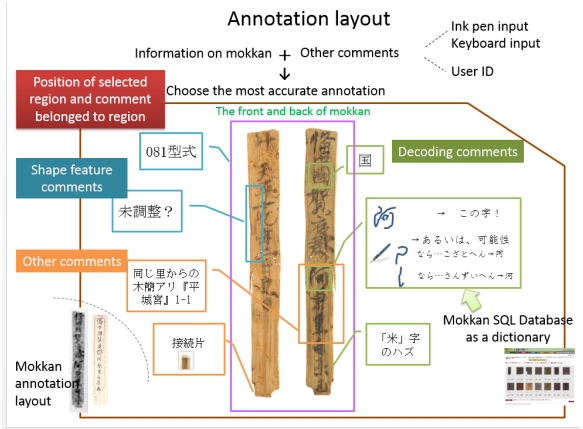

Fig. 4: Annotation layout from the suggestion of NNRICP.

1) Decoding comment (green): users can input their interpretations, opinions and considerations on recognizable and unrecognisable characters.
2) Shape feature comment (blue): mokkan piece properties such as model, width, shape and height.

3) Other comment (orange): characteristics on a mokkan piece such as marks, corners, edges, and likeliness to be connected to other pieces, and so on.

Each comment will be accompanied with user ID, comment type. With decoding comments, the system will provide the following options to help users interpret characters. On the other hand, since the second and third types of comments do not include text, users can input by mouse or keyboard and get hints of comments at the same position but in another annotation layout.

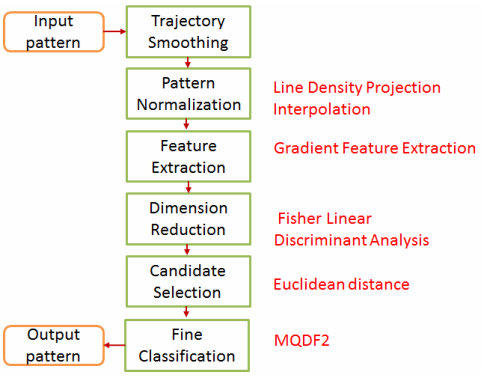

Fig. 5: The steps of handwriting recognition.

1) Using our off-line handwriting recognizer to get a list of candidates on the selected area automatically. Even if a character image is badly damaged, it can produce candidates and hints for archaeologists. Off-line recognition degrades its performance when patterns are damaged but can produce candidates even if patterns are badly damaged while people can read damaged patterns but has no idea for badly damaged patterns. The steps of handwriting recognition is shown in Figure 5.
2) Referring the previously annotated comments by other colleagues at the same position on a mokkan piece layout or finding with radical, number of strokes, Unicode in mokkan SQL database.
3) Writing annotations and employing our on-line recognizer. Otherwise, employing another option to input comments using mouse and keyboard.
4) Drawing characters with a pen or a finger on a touch pad screen. The information of ink comments will be saved to XML input file and loaded again each time the system is started.

D. Mokkan piece clustering

After adding annotations, the system allows users to classify pieces of mokkan into groups, based on three types of annotation: decoding comments, shape feature comments and other comments as shown in Figure 6. In the paper [6], we proposed a solution to cluster mokkan pieces, including two steps: an image grouping using color features and an image reassembling using local tangent and curvature functions of the fragment contours. In the first step, we extracted color features by three functions: mean (1), standard deviation (2), and skewness (3). Mean is the average color value in the

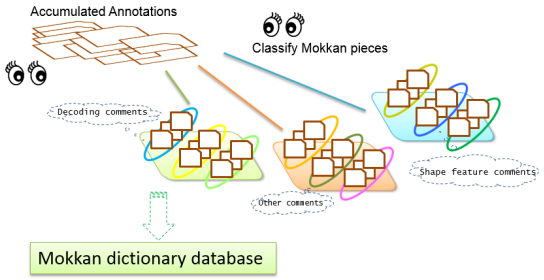

Fig. 6: Cluster mokkans from the annotated information.

image. The standard deviation is used to measure confidence of the distribution. And the skewness is a measure of the degree of asymmetric in the distribution.

Mean:

$$\mu_i = \frac{1}{W * H} \sum_{x=1}^{W} \sum_{y=1}^{H} p_{xy} \tag{1}$$

Standard deviation:

$$\sigma_i = \sqrt{\frac{1}{W * H} \sum_{x=1}^{W} \sum_{y=1}^{H} (p_{xy} - \mu_i)^2} \tag{2}$$

Skewness:

$$\gamma_i = \sqrt[3]{\frac{1}{W * H} \sum_{x=1}^{W} \sum_{y=1}^{H} (p_{xy} - \mu_i)^3} \tag{3}$$

where W, H is the width and heigh of image respectively, p_{xy} is the value of color at x,y pixel.

Then we used the k-Means algorithm and the LBG algorithm to classify mokkan pieces. In the second step, we proposed a coarse-to-fine curve matching method based on the boundary curve of mokkan to finely classify mokkan. This worked for artificially broken pieces but not for real mokkan pieces. Grouping mokkan pieces based on color is effective

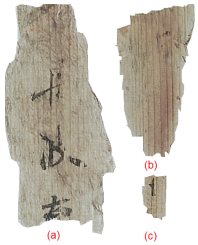

Fig. 7: The grouping result in the paper [6]

because mokkan peices, buried in the same grid, often have the similar color due to the same impact of environment and climate. Deriving from the previously archived result, firstly we will also group mokkan pieces based on color features, but with some improvement. In the previously proposed method, because we calculated color features on the whole of mokkan image, the result was misleading if there was ink on the image such as shown in Figure 7 (it is clear that mokkan pieces a, b, c in Figure 7 have the different colors, so they should be classified into the different groups). Instead of calculating color features on the whole of mokkan image, we will apply

the mean filter to equalize the whole image (this step will be effective because of wood vein, and ink marks). Then, we will take the most typical window of adjusted image with the fixed dimension, map to the correspondent part of original image conversely, and extract color features from the window as shown in Figure 8. After having groups of

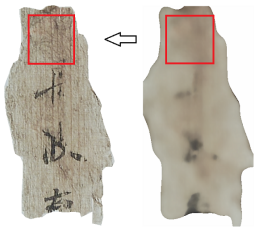

Fig. 8: The proposal to extract color feature of mokkan pieces.

mokkan pieces from the first step, we analyze the semantic of comments to finely classify mokkan pieces. For example, from the decoded characters of a mokkan piece, we can categorize it into the group of other mokkan pieces which contain the same decoded characters. The grouping of mokkan pieces based on the semantic of annotations can be either performed automatically by the system as the above example or by archaeologists. For instance, once an archaeologist sees the shape feature describing of a mokkan piece and then finds the similar description on the other piece, he/she can group them into the same group. In any manner, the classification by the semantic of annotations is extremely difficult and needs more research in the future such as using neural networks and other methods to process big data.

III. The Proposed GUI of System

MokkAnnotator II is being developed in C# language with the Microsoft SQL server. It is constituted of five components: menu and tool bar (A), project explorer (B), annotation layout (C), property window (D), mokkan data grid dictionary (E). The GUI is built as MDI (Multiple Document Interface) in which multiple annotation layouts can be opened in separate tabs simultaneously. Each component in GUI is a dock item, so that users can change its position easily.

(A) Menu and tool bar: allows users to open, save, close projects and provide buttons to apply image processing and to add annotations.

(B) Project explorer: allows users to open multiple projects (mokkan pieces) simultaneously. Mokkan projects are organized into folders which contain some annotation layouts created by users. Basically, each layout is a type of image to which different methods are applied as section II-B and users can add new mokkan layouts freely.

(C) Annotation layout: allows users to add three types of comment using buttons on tool bar.

(D) Property window: displays the information of a mokkan piece, users, and annotation layouts.

(E) Mokkan data grid dictionary: displays annotated comments stored in SQL database as a mokkan dictionary. Users can find characters in the database with radical, Unicode, and number of strokes.

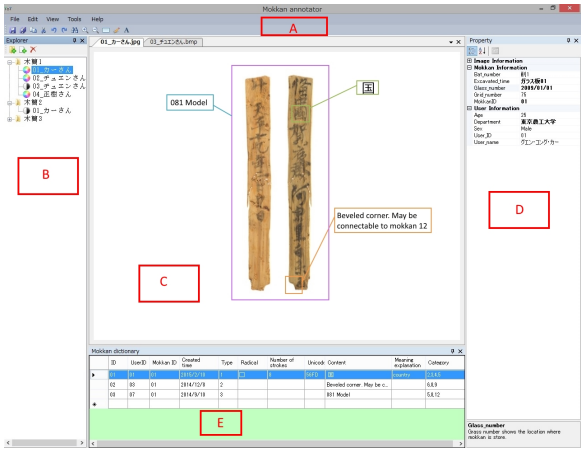

Fig. 9: The proposed GUI of Mokkannotator II.

IV. Conclusion

This paper has presented a system for archaeologists to enhance images for mokkan interpretation, annotate mokkan pieces, share annotations among them and cluster mokkan pieces based on annotations, shape and color. For techniques to enhance images for mokkan interpretation, we use methods proposed in the paper [3] while for supporting users to annotate, we exert the most advanced technologies performed at our laboratory [5]. We also improve the method proposed in the paper [6] to obtain the better grouping result. Then we analyse the semantic of annotations and then categorize mokkan pieces into groups. Although the work is difficult, we believe it will open the new research direction for grouping mokkan pieces.

Acknowledgment

This work is being supported by the Grant-in-Aid for Scientific Research (S)-20222002. The authors would like to thank the archeologists in NCPRI for data, material, suggestions and introducing the current work, mokkan repository to us.

References

[1] 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.

[2] Phan, T. V, Baba, H., Watanabe, A., and Nakagawa, M: MokkAnnotator - An Annotation Tool to Accumulate and Organize Mokkans. In Proceedings of the 15th International Graphonomics Society Conference (IGS2011) (Cancun, Mexico, Jun. 12-15, 2011), pp.152-155.

[3] Takakura, J., Kitadai, A., Nakagawa, M., Baba, H., and Watanabe, A: Techniques to Enhance Images for Mokkan Interpretation. In Proceedings of the 12th International Conference on Frontiers in Handwriting Recognition (ICFHR2010) (Kolkata, India, Nov. 16-18, 2010), pp.358-362.

[4] Akihito Kitadai, Kei Saito, Daisuke Hachiya, Masaki Nakagawa, Hajime Baba and Akihiro Watanabe: Design and Prototype of a Support System for Archeologists to Decode Scripts on Mokkan. Proc. 13th Conference of the International Graphonomics Society (IGS), Salerno, Italy (2005.6), pp.54-58.

[5] Phan, T. V; JinFeng Gao ; Bilan Zhu ; Nakagawa, M: Effects of Line Densities on Nonlinear Normal-ization for Online Handwritten Japanese Character Recognition. Document Analysis and Recognition (ICDAR), 2011 International Conference on IEEE, pp.834 - 838.

[6] Van Phan, T., Baba, H., Watanabe, A., & Nakagawa, M: A re-assembling scheme of fragmented Mokkan images. In Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing ACM, pp. 22-28.