Similarity Evaluation and Shape Feature Extraction for Character Pattern Retrieval to Support Reading Historical Documents

Akihito KITADAI
J. F. Oberlin University
Tokiwa-machi 3758, Machida, Tokyo, Japan
e-mail: a.kitadai@gmail.com

Masaki NAKAGAWA
Tokyo University of Agriculture & Technology
Naka-machi 2-24-16 Koganei, Tokyo, Japan
e-mail: nakagawa@tuat.ac.jp

Hajime BABA and Akihiro WATANABE
Nara National Research Institute for Cultural Properties
Nijo-cho 2-9-1, Nara, Japan
e-mail: {hajime, akihiro}@nabunken.go.jp

***Abstract*— We have many historical documents written in over 1,000 years ago. Shape features of character patterns on the documents are unstable or missing because most of the documents have been stained and degraded deeply. Digital archives of the documents with accurate character pattern retrieval methods are helpful for archaeologists and historians. In this paper, we propose a similarity evaluation method for character patterns with missing shape parts. It collaboratively works with non-linear normalization for such patterns, and modifies the templates for each trial of the retrieval efficiently. In the experiences using 4,911 *Kanji* (Chinese origin) character patterns from the Japanese historical documents called *mokkan*s, the method shows improvements of the retrieval accuracy. Also, we present a simple implementation of gradient feature extraction to compare the chaincode feature with the gradient feature in the retrieval. As the result, the gradient feature works better than the chaincode feature.**

Keywords-character pattern retrieval, historical documents, mokkan, gradient feature

I. INTRODUCTION

Historical documents are important properties when we consider ancient cultures and human activities. Many archaeologists and historians are trying to analyze and decode the documents.

More than 1,000 years ago, Japanese people at the time produced many documents by using wooden tablets, brushes and ink. We call each of them "*mokkan*" that means wooden document (Fig. 1). We have more than 320,000 historical *mokkans* excavated from Japanese old ruins. Especially, more than 180,000 of them are from the ruin *Heijo* in Nara prefecture. The ruin was the capital of Japan from A.D. 710 to 784. The analysis and decoding results of the *mokkans* provide much precious information for the study areas of archaeology and history: money or material flows, place and human names, social structures and so on. However, since most of the *mokkans* are too old and have been stained and scratched in under the ground of the ruins, to read such degraded *mokkans* is difficult even for archaeologists and historians.

Figure 1. Historical *mokkans*.

Recently, several researches of computer science have proposed image processing methods to extract character patterns on degraded historical documents [1-3]. Also, digital archives of historical documents are constructed in study areas of archaeology and history. Character pattern retrieval (CPR) enhances these research products by accepting the extracted character patterns as the keys and providing their similar character patterns from the archives. Additionally, the CPR constructs web links to the archived documents that contain the similar patterns. The results of retrieval become the valuable references to support reading the degraded historical documents including the *mokkans* [4].

978-0-7695-4661-2/12 $26.00 © 2012 IEEE
DOI 10.1109/DAS.2012.80

One of the serious problems for the CPR is missing shape parts of character patterns. On the historical *mokkans*, decayed and discolored surfaces of the wooden tablets and tarnished ink generate lots of missing shape parts in the character patterns. Therefore, we need robust techniques of similarity evaluation for the CPR. Another problem is unstable shape features of character patterns. Even if we employ accurate image processing, the patterns contain noises coming from degraded documents.

II. NON-LINIEAR NORMALIZATION METHOD

In this section, we introduce our interactive non-linear normalization method that we have proposed [4].

Non-linear normalization using histograms of shape features in *X*- and *Y*- coordinates performs well in character pattern recognition [5]. However, deformation by the normalization is too much for the keys of character pattern retrieval with missing shape parts. Fig. 2 shows the problem: $h(x)$ and $h(y)$ are the histograms, $a(x)$ and $a(y)$ are the accumulative functions of $h(x)$ and $h(y)$, $a(x')$ and $a(y')$ are the linearized $a(x)$ and $a(y)$.

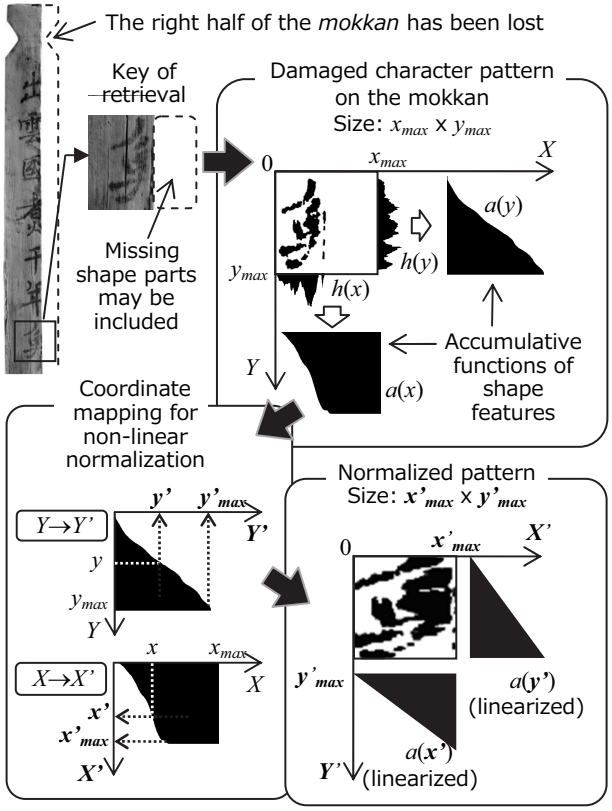

Figure 2. Problems of non-linear normalization.

To manage the deformation, our normalization method employs the gray-zones painted by the archaeologists and historians with digital pointing devices. We show the flow of the method in Fig. 3: $h_b(x)$ and $h_b(y)$ are the histograms of the pattern with a blacken gray-zone, $a_b(x)$ and $a_b(y)$ are the accumulative functions of the $h_b(x)$ and $h_b(y)$. In this method, we linearize the weighted averages of the accumulative function in each direction: $a_{w\_ave}(x)$ and $a_{w\_ave}(y)$.

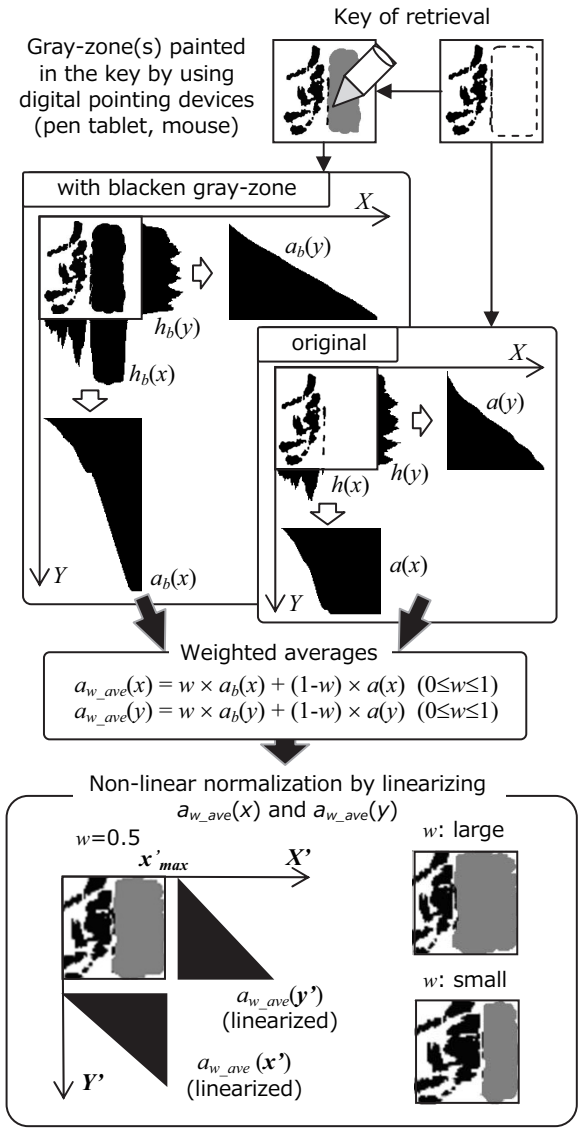

Figure 3. Interactive non-linear normalization method.

The weight w in Fig. 3 can take the value from 0 to 1 to control the deformation. A small value of w assumes a small amount of missing shape parts in the gray-zone, while a large value of w means a large amount of them.

The above method with line density equalization is working on our CPR system for historical documents [6]. Also, we have extended the method in order to vary the value of w with the location of each gray-zone in a key.

Fig. 4 shows an example of character pattern retrieval by using our CPR system.

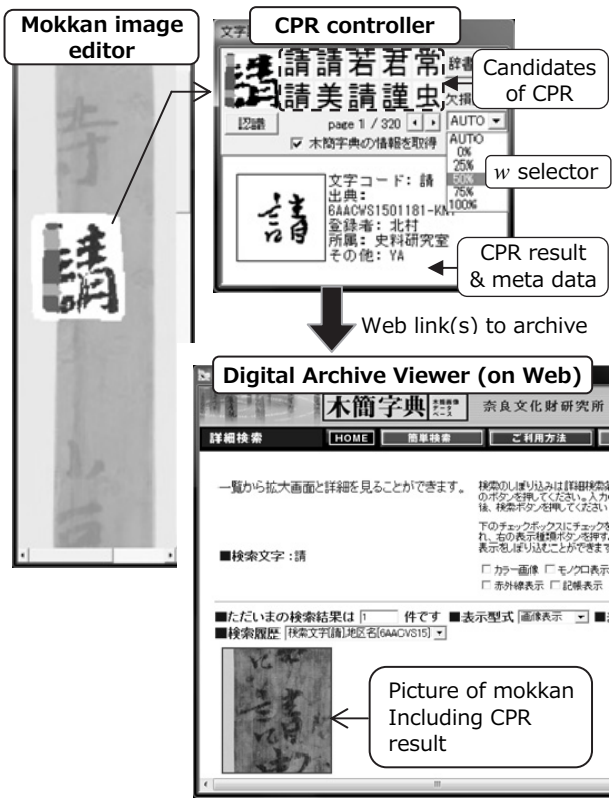

Figure 4. Our CPR system Working with *Mokkan* Digital Archive.

III. SIMILARITY EVALUATION FOR CPR

A. Creating Feature Matrix

In our CPR method, we employ chaincode features as the shape features of character patterns. By scanning the pixels in the non-linear normalized character patterns with the following 3×3 pixel-patterns (Fig. 5), we obtain 4-directional chaincode features. The pixel-patterns enable us to perform the contour tracking and direction partitioning for the chaincode extraction by a single pass of the scanning [5].

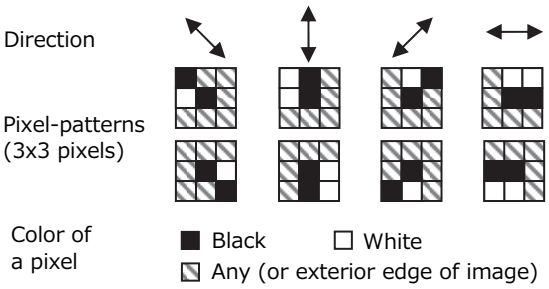

Figure 5. Pixel-patterns for 4-directional chaincode features.

In the sampling process of our CPR method, we accumulate the shape features with multiplying Gaussian functions (Fig. 6). Each of the Gaussian functions has a unique center point that is an intersection point of the grid, and amplifies the features close to the point. In this paper, we present one of the Gaussian functions as $G(h, v)$. The (h, v) is the center point in which h, v x shows the column number and y shows the row number of the grid.

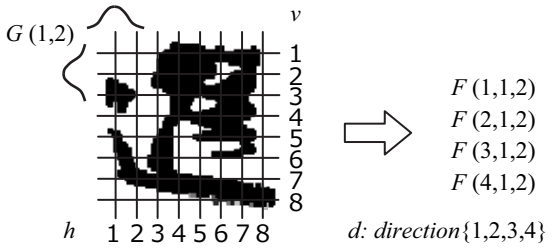

Figure 6. Sampled features: $F(d, h, w)$ with a Gaussian function $G(1, 2)$.

We use the 8×8 equally-separated grid for the Gaussian functions to obtain 64 accumulated shape features for each direction. We present each of the accumulated shape features $F(d, h, v)$. Finally, the dimension of feature matrix becomes 4×64 for each character pattern.

The feature matrices of the character pattern images in the archives of the *mokkans* become the templates of our CPR. For the similarity evaluation between each of the templates and a key, we employ the negative value of the city block distance. This measure has shown good performance in our previous researches.

B. Problem in Similarity Evaluation

Even if the gray-zones correct the X- and Y- coordinates of shape features in the normalized key, the missing shape features are not sampled from the gray-zone. It produces unjust distance between the key and the genuine template(s) of retrieval. The distance will be absorbed if we also apply the gray-zone to all character pattern images in the archives and reconstruct all templates from the images. However, it is not practical because this reconstruction of the templates is necessary in each trial of the retrieval.

We have proposed an alternative method [7]. In this paper, we call it averaged feature (AF) method. The AF method does not reconstruct the templates. Instead, it obtains the averaged shape features outside of the gray-zone in the normalized key, and injects it into the gray-zone.

Since the remaining and missing shape parts in a same character pattern were written by the same person with the same brush in the same environment, the averaged shape features work better than no shape features. However, we need more accurate method than the AF method that considers the local shape differences in the pattern.

IV. SOLUTION FOR SIMILARITY EVALUATION

In this paper, we propose the other method to absorb the unjust distance efficiently. In this paper, we call it template modification (TM) method. The TM method does not reconstruct the feature matrices of the templates or inject any shape features into the gray-zone. In each calculation for similarity evaluation, the TM method just reduces each $F(d, x, y)$ in the feature matrices of the templates by considering the gray-zone in the normalized key.

For each pixel in the normalized key with gray-zone, the TM method gives a score s_{gray} shown in (1).

$$s_{gray} = \begin{cases} a \ (\textit{for pixels in gray-zone}) \\ 0 \ (\textit{for pixels not in gray-zone}) \end{cases} \quad (1)$$

The value a is constant and not equal to 0. Also, the TM method gives the other score $s = a$ to every pixel in the normalized key.

As same as the sampling process of shape features, the TM method accumulates the s_{gray} and s with the same Gaussian functions $G(h, v)$. We present each accumulated result as $S_{gray}(h, v)$ and $S(h, v)$. Finally, the TM method replaces the $F(d, h, v)$ in the feature matrix of the templates by the $F'(d, h, v)$ as presented in (2).

$$F'(d, h, v) = F(d, h, v) \times \{1 - S_{gray}(h, v) / S(h, v)\} \quad (2)$$

Equation (1) and (2) do not contain the weight w in Fig.3 explicitly. The value reflected in the size of gray-zone in the normalized key is considered by the $S_{gray}(h, v)$.

V. Experimental Results

To show the improvements of the TM method, we employed 4,911 binary images of *Kanji* character pattern from a digital archive of the *mokkans* [8]. Each of them is barely legible for human readers (Fig. 7). Also, we made 10 mask images to generate missing shape parts or gray-zones on the character pattern images artificially (Fig. 8). Such managed missing shape parts and gray-zones are suitable for quantitative evaluations of CPR. Additionally, we applied uniform scaling to the character pattern images. This scaling fits the long side of character pattern boundary box to 64 pixels since the size of each mask image is 64×64 pixels.

Figure 7. Binary Images of Character Patterns from *Mokkans*.

We used leave-one-out cross validation method in the following experiments of character pattern retrieval [9]. Every character pattern image except a key became a template in each trial of the retrieval. Also, we set criteria for the trials: the trial was "hit" if the character codes of "top 10 similar templates" to the key contained the code of the key.

The hit rate without the mask images was 81.7% (4,012/4,911). Also, the hit-rate was 44.2% (21,667/49,016) when we used the mask images with the w=0 (missing shape parts). Since all black pixels in a key were completely covered by a mask image, we did not employ 94 trials in the latter experiment (4,911×10-94=49,016). The missing shape parts degraded the accuracy of the character pattern retrieval.

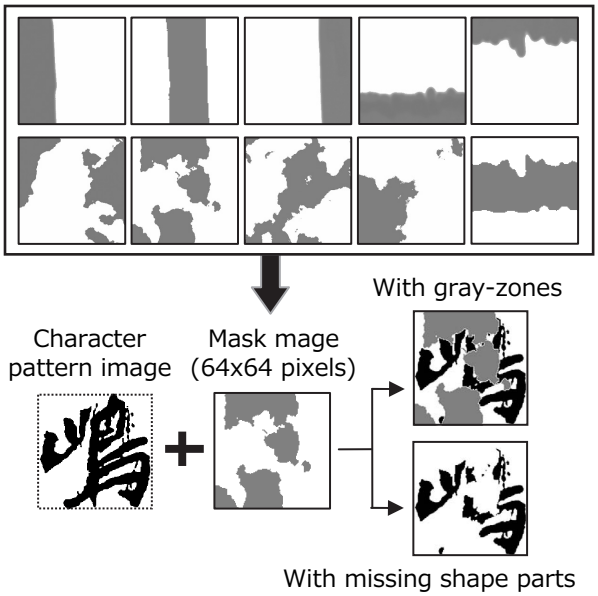

Figure 8. How to Use Mask Images.

The table I shows the hit rates with the mask images and the w=0.5. The table II shows another results with the mask images when we employ the optimal value of w among 0/0.25/0.5/0.75/1 for each trial.

TABLE I. Hit Rates Using The Mask Images (w=0.5)

	Hit Rates
AF method	53.9% (26,428/49,016)
TM method	**60.4**% (29,608/49,016)

TABLE II. Hit Rates Using The Mask Images (w: Optimal)

	Hit Rates
AF method	66.1% (32,394/49,016)
TM method	**72.7**% (35,639/49,016)

Even if the AF method regained the retrieval accuracy, the TM method performed better than it. Utilizing the gray-zones more effectively is our future work.

VI. Discussion to Use Gradient Feature

Color or gray-scale images of the *mokkans* are too noisy to employ the original gradient features for character pattern retrieval. This is one of the reasons why we use binary images of character patterns in our retrieval method. Even so, as shown in Fig. 7, many noises are still remaining around the contour pixels of the character patterns.

Some recent researches of character pattern recognition have proposed the methods of artificial gradient feature extraction from binary images [10]. The methods containing soft decision of the stroke-orientations extract stable features and fit together with machine learning methods well. We still have problems with applying the machine learning methods in consideration of the missing shape parts and gray-zones. However, the stable shape features produced by the soft decision may improve the accuracy of CPR for historical documents.

To consider the effect of the soft decision, we made a small modification of the feature extraction process. In this section, we present the *X*- and *Y*- coordinates of a contour pixel $p_c'=(x_c', y_c')$ in a non-linear normalized character pattern image. The contour tracking is performed by using the pixel-patterns in Fig. 5.

In the first step of the modified process, we obtain the $p_c=(x_c, y_c)$ that is the original p_c' of the character pattern image before the non-liner normalization. The long side of the character pattern boundary box has been fit to 64 pixels. In the next step, we apply the 3×3 weighted average filter to every pixel of the image including p_c. To create nearly-linear gradients around the p_c, we repeat the filter application thrice. Then we calculate the 4-directional gradient features *g* as shown in Fig. 9.

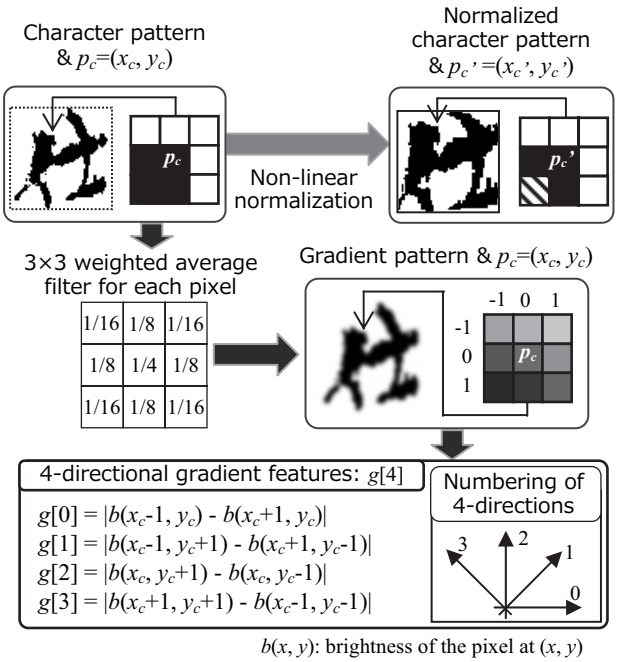

Figure 9. Process to Obtain Gradient Features.

Finally, we remove the minimum element of the gradient features as noise-reduction, and assign the other three elements to p_c' instead of the chaincode features. Table III shows the result of the modification.

TABLE III. HIT RATES WITH GRADIENT FEATURE

	Hit Rates
Without Mask Images	82.5% (4,049/4,911)
TM method (w=0.5)	61.8% (30,307/49,016)
TM method (w: optimal)	74.8% (36,639/49,016)

Since the modification replaces only the shape features, we consider that the soft decision of the stroke orientation is better for CPR with missing shape features. More improvements and considerations of shape feature extraction are our future work.

VII. CONCLUSION

In this paper, we present an improvement of similarity evaluation and feature extraction methods for the CPR to support reading historical documents. More improvements to recover the degradation of the retrieval accuracy caused by the missing shape parts and to utilize the gradient features are our future work.

ACKNOWLEDGMENT

This work was supported by the Grant-in-Aid for Scientific Research (S)-20222002 and Young Scientists (B)-22720239.

REFERENCES

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

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

[3] J. Takakura, A. Kitadai, M. Nakagawa, H. Baba and A. Watanabe, "Techniques to Enhance Images for Mokkan Interpretation," Proc. 12th ICFHR, Vol.1, No.1, pp.358-362, Kolkata, India, Novenber 2010.

[4] A. Kitadai, M. Nakagawa, H. Baba and A. Watanabe, "A Combining Method of Non-linear Normalization to Support Reading Damaged Character Pattern on Historical Documents," Proc. 13th IGS, Vol1. pp.2-5, Dijon, France, September 2009.

[5] C.L. Liu, Y.J. Liu, R.W. Dai, "Multiresolution statistical and structural feature extraction for handwritten numeral recognition," Pre-Proc. 5th IWHFR, pp. 61-66, Colchester, England, August 1996.

[6] H. Yamada, K. Yamamoto and T. Sato, "A Nonlinear Normalization Method for Handprinted Kanji Character Recognition ---Line Density Equalization---," Proc. 9th ICPR, Vol. 1, pp.172-175, Roma, Italy, .

[7] M. Nakagawa, K. Saito, A. Kitadai, J. Tokuno, H. Baba, A. Watanabe, "Damaged character pattern recognition on wooden tablets excavated from the Heijyo palace site," Proc. 10th IWFHR, La Baule, France, Vol.I, pp.533-538, October 2006.

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

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

[10] C-L. Liu, "Handwritten Chinese Character Recognition: Effects of Shape Normalization and Feature Extraction," Lecture Notes in Computer Science, Vol. 4768/2008, pp.104-128, 2008.