Writer Identification for Offline Handwritten Kanji without using Character Recognition Features

Ayumu Soma
Teikyo University
Graduate School of Science & Engineering
1-1 Toyosatodai, Utsunomiya, Tochigi, Japan
12M105@uccl.teikyo-u.ac.jp

Masayuki Arai
Teikyo University
Graduate School of Science & Engineering
1-1 Toyosatodai, Utsunomiya, Tochigi, Japan
arai@ics.teikyo-u.ac.jp

***Abstract*— Most research on writer identification in the case of offline handwritten Kanji characters uses character recognition features. In this paper, we assume the following is representative of the writers' style: the start point, the end point, the angle of each stroke composing a Kanji character, and the size and position of the Kanji character. The experimental results show that the identification rates are 95.2% without rejection and 99.6% with 10% rejection for four Kanji characters written by one hundred writers.**

Keywords- writer identification; handwritten characters; offline characters; Kanji

I. INTRODUCTION

Writer identification based on scanned images of handwritten characters is a useful biometric modality with applications in forensic and historical document analysis. Research on writer identification that uses online characters is widespread, even though online characters lack form for conveying dynamic information. However, very little research on offline characters has been reported [1][2].

In the research on offline Kanji, which consists of logographic Chinese characters adopted in Japanese writing, character writer identification is also not common. Most identification methods use features developed for character recognition, such as the weighted direction index histogram feature [3], the local direction contributivity feature [4], and the directional element feature [4]. However, we assume these features are not efficient for writer identification because they do not reflect the writers' handwriting style. For this reason, in this paper, we employ the following features for writer identification: the start point, the end point, the angle of each stroke that composes a Kanji character, the character size, the center point of gravity, and the position of the character.

II. FEATURES EXTRACTED FROM EACH STROKE

Tokiwa et al. reported that the start point and the end point of each stroke of Kanji characters are efficient features for writer identification [5]. However, in their work, the strokes were extracted manually. Therefore, we extract them automatically and employ the same features and angles of the stroke.

Before feature extraction, Kanji character images are preprocessed as follows: images are binarized and then thinned so that each stroke is one pixel width. After preprocessing, each stroke comprising the character is extracted. The following explains the procedure for extracting the horizontal strokes shown in Fig. 1.

(1) Scan vertically from point H1 and find the first black pixel H2, as shown in Fig. 1. Then, pixel H2 is the start point of a stroke.
(2) If one of three neighboring pixels—first, the right pixel; second, the upper right pixel; and third, the lower right pixel—is black, the scan proceeds to that pixel.
(3) Repeat step (2) until the three neighboring pixels are all white or the scan reaches the end of the image.
(4) If stroke candidates are found, the longest one is selected. If some candidates have the same length, the one located in the upper position is selected. Therefore, in the case of Fig. 1, the end point of the stroke whose start point is H2 is determined to be H3. Here, the pixels already scanned are not selected as the start and end points.

The procedure for extracting vertical strokes is almost the same as that for the horizontal strokes above, except the three neighboring pixels in step (2) are replaced with the lower pixel, the lower left pixel, and the lower right pixel, respectively. Therefore, for the example given in Fig. 1, the stroke having the start point H4 and the end point H5 is extracted.

Strokes whose lengths are less than three pixels are not extracted because they are likely to be noise. Next, the angle of the stroke is determined from a straight line linking the start point to the end point. Here two strokes, which have following conditions, are be identified as the same stroke and are merged: the difference between the angle of the two strokes is less than or equal to ten degrees, and the position of the start point or the end point is within eight pixels of another stroke.

Finally, for each stroke, we obtain five features: the x- and y-coordinates of the start point, the x- and y-coordinates of the end point, and the angle of the stroke.

Published by Atlantis Press, Paris, France.

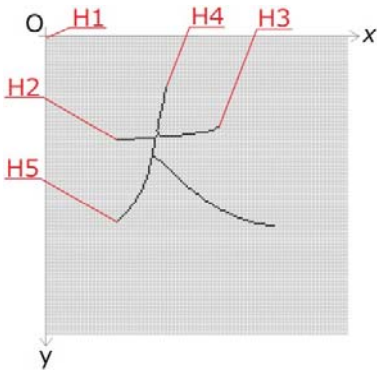

Figure 1. Example of extracting strokes from a character.

III. SIZE AND POSITION OF KANJI CHARACTERS

In this section we describe the method to extract the size and position of Kanji characters by referring to Fig. 2. The features, which do not need preprocessing, are as follows:

(1) the x- and y-coordinates of the upper left point of the inscribed rectangle for the character (Fig. 2, p1)
(2) the x- and y-coordinates of the lower right point of the inscribed rectangle for the character (Fig. 2, p2)
(3) the x- and y-coordinates of the center point of gravity where all the black pixels are concentrated (Fig. 2. g1)
(4) the character width plus the character height (Fig. 2, w2+h2)
(5) the character width divided by the character height (Fig. 2, w2/h2)
(6) the character width divided by the frame width (Fig. 2, w2/w1)
(7) the character height divided by the frame height (Fig. 2, h2/h1)

Therefore, these features consist of ten elements.

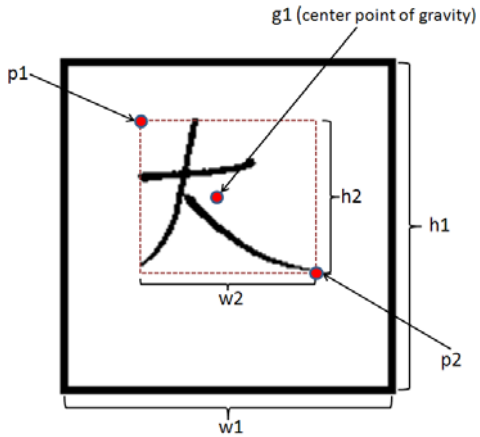

Figure 2. An example of the global features.

IV. EXPERIMENTS AND DISCUSSION

A. Experimental Conditions

In the present study, we used an offline Kanji character database [6]. The database contains one hundred Kanji characters, each written by one hundred writers, and fifty samples of each Kanji character for a given writer. Therefore, the database registers 500,000 samples. One character image has 160 160 pixels.

First, we selected the four Kanji characters, “御”, “前”, “崎” and “市”, which together represent a city name “御前崎市” (Omaezaki-shi). The reasons for selecting these characters are as follows. First, in general, some kinds of characters can be used in the research field on writer identification; second, for research on text-independent writer identification, as in this paper, more broadly used words are suitable for experiments. However, very few general words can be constructed using only one hundred Kanji characters.

On the assumption that the Kanji characters have already been recognized correctly, we conducted an experiment with each kind of Kanji character. For each experiment, 5,000 samples (100 writers 50 samples) were used in the tests using leave-one-out cross validation; that is, one sample was used for validation, while the other 4,999 samples were used for the dictionary.

The method used to compare the validation sample and the dictionary sample and to identify the writer is as follows:

(1) Calculate the Euclidian distances between each five-dimensional vector (the features extracted from a stroke) of the validation sample and of the dictionary sample, and then choose the shortest distance d1 from all combinations of feature vectors of two samples. If the number of extracted strokes is different between the two samples, the number of validation sample strokes is used for the distance calculation. In the case that the number of the dictionary sample strokes is less than that for the validation sample, the same stroke or strokes of the dictionary sample are used more than once for distance calculation.
(2) Calculate the Euclidian distances d2 between each ten-dimensional vector (the features for the character size and position) of the validation sample and of the dictionary sample.
(3) Repeat steps (1) and (2) the same number of times as the number of Kanji characters.
(4) Rank the writers of the dictionary samples in ascending distance d1+d2 order for each Kanji character.
(5) Rank the writers of the dictionary samples in ascending distance d1+d2 order for each Kanji character. The first writer which is being met every Kanji character is recognized as the writer of the validation sample.

B. Experimental Results and Discussion

Fig. 3 depicts the identification rate as a function of the number of writers. The identification rates for ten writers, fifty writers, and one hundred writers are 99.60%, 97.04%,

and 95.22%, respectively. The rate decreases linearly with the number of writers.

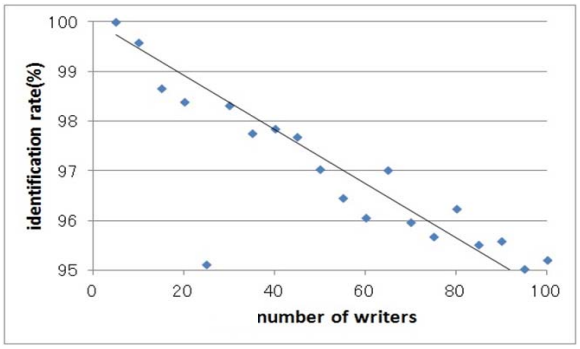

Figure 3. Identification rate vs. number of writers.

Fig. 4 depicts the identification rate using one hundred writers when we employed the rejection. Fig. 4 shows that the identification rate is 99.6% with 10% rejection.

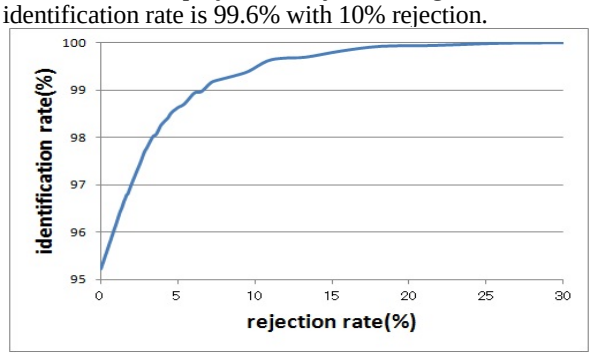

Figure 4. Identification rate vs. Rejection rate.

Some papers report that the weighted direction index histogram (WDIH) can achieve high recognition rates for writer identification [3]. Therefore, we compared the WDIH method with the method proposed in the present paper. Fig. 5 depicts the identification rate using the WDIH. It can be seen that the identification rates in Fig. 5 are almost the same as those in Fig. 3. Therefore, wit can be concluded that the proposed feature-identification method is effective for writer identification.

V. CONCLUSION

Most research on writer identification using offline handwritten Kanji characters employs character recognition features. In this paper, we employ the following features: the start point, the end point, the angle of each stroke composing a Kanji character, and the size and position of the Kanji character. The experimental results show that the identification rates are 95.2% without rejection and 99.6% with 10% rejection for four Kanji characters written by one hundred writers.

Future work will consider other features to improve the identification rate.

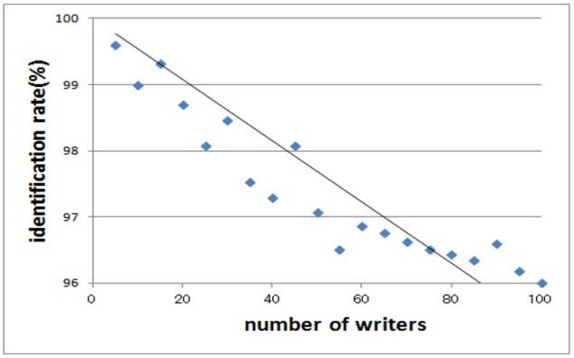

Figure 5. Identification rate vs. number of writers using the weighted direction index histogram.

ACKNOWLEDGMENT

The authors would like to thank Mr. Kiichi Misaki for lending us the offline Kanji database.

REFERENCES

[1] M. Bulacu, and L. Schomaker, “Text-independent Writer Identification and Verification Using Textural and Allographic Features,” IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 29, no. 4, pp. 701-717, 2007.

[2] M. K. Kalera, S. Srihari, and A. Xu, “Offline Signature Verification and Identification Using Distance Statistics,” International Journal of Pattern Recognition and Artificial Intelligence, vol. 18, no. 7, pp. 1339-1360, 2004.

[3] M. Umeda, T. Miyoshi, and K. Misaki, “Writer Identification and Verification Using Autoassociative Neural Networks,” Trans. of the Institute of Electrical Engineers of Japan. C, A publication of the Electronics, Information and System Society, vol. 122, no. 11, pp. 1869-1875, 2002 (in Japanese).

[4] K. Misaki, and M. Umeda, “Handwriter Identification Using Quantitative Features Extracted from Character Patterns,” Trans. of Japanese Association of Forensic Science and Technology, vol. 2, no. 2, pp. 71-77, 1997 (in Japanese).

[5] K. Tokiwa, K. Fukue, and Y. Matsumae, “Personal Identification by Handwriting Based on Only Layout Pattern of Strokes Restraining Intra-individual Variation,” Trans. of the Institute of Image Electronics Engineers of Japan, vol. 40, no. 4, pp. 660-670, 2011 (in Japanese).

[6] K. Misaki, D. Honjou, and M. Umeda, “Database of Handwritten Characters for Writer Recognition Study and Software for Display and Analysis,” Trans. of Japanese Association of Forensic Science and Technology, vol. 7, no. 1, pp. 71-81, 2002 (in Japanese)