Selection of a Similar Writer for Personal Handwritten Character Recognition

Takuya Kimura[1], Masahiro Hattori[2], Shinji Tsuruoka[1],
Hiroharu Kawanaka[2], Haruhiko Takase[2] and Yasuji Miyake[2]

[1] Graduate School of Regional Innovation Studies, Mie University, Mie, Japan
[2] Graduate School of Engineering, Mie University, Mie, Japan

E-mail: kimura@ip.elec.mie-u.ac.jp

Abstract— **It becomes the diffusion period of a tablet (keyboard less) computer for electronic book and electronic publishing, and every business elite will bring a personal tablet computer. The usage of personal tablet computer includes taking memo using the touch panel display. The taking memo using the touch panel display is written by a finger or a pen. The memo includes a figure and a character written by his finger or pen. The user desires the handwritten character recognition. The useful handwritten character recognition software, however, has not developed for Japanese characters. We will develop the personal handwritten character recognition system in high accuracy using the tablet can use the trajectory of his finger or pen. In this paper, we describe the effectiveness of personal character recognition and a selection method of a similar writer in personal character recognition. We proposed two new recognition methods: "similar mean dictionary (SMD)" and "similar feature space dictionary (SFSD)" to improve the personal character recognition. These dictionaries use only one character, and they can estimate the personality of writing style using characters written by a similar writer. The similar feature space dictionary improved the character recognition rate to 91.0 % which obtains the similar accuracy with personal dictionary. It is confirmed that only one character written by a user (specific writer) is very effective for handwritten character recognition.**

Keywords— **Handwriting, Hiragana recognition, Character recognition, Handwriting character recognition, Personal tablet computer**

I. Introduction

In recent years, a personal tablet computer such as iPad, kindle and so no is becoming popular as an electronic book and an electronic publish is spreading. The most business elite will bring the personal tablet computer. On the personal tablet computer, a character is written by a finger or a pen on the touch panel display, and business elites will require the character recognition with personal tablet computer to compress the information. Handwritten character recognition software is used for English document, however, recognition accuracy for Japanese characters is not adequate. The problem of Japanese handwritten character recognition is increasing for popularization.

To improve the accuracy of character recognition, we could use a personal recognition dictionary which is made of a user's (specific writer's) characters. The problem of personal recognition dictionary needs many characters handwritten by a specific writer [1, 2].

In this paper, we propose two types of dictionary generating methods for personal recognition. The proposed methods use the selecting a similar writer for writing style by only one handwritten character. The first type dictionary "similar mean dictionary (SMD)" uses the mean feature vector of the similar writer as the personal dictionary. The second type dictionary "similar feature space dictionary (SFSD)" uses the mean feature vector and the covariance matrix of the similar writer. We used 71 kinds of handwritten "HIRGANA" characters for each writer to confirm the effect of character recognition.

II. Personal recognition dictionary

The personal recognition dictionary has not used generally as an application using the personal characters written by a specified writer. Usually, the writer who takes the memo in personal tablet computer is only one person. The character form has the personal writing habit (Fig.1).

The usage of the character form written by a specific writer is effective to improve the recognition accuracy of character recognition.

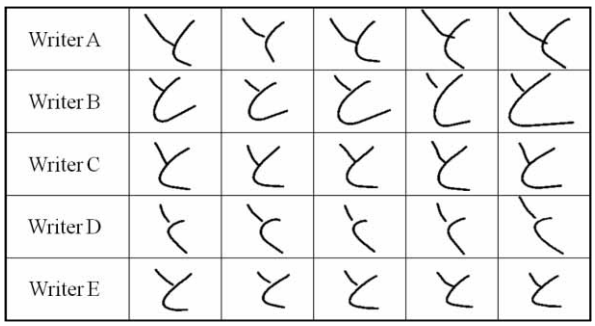
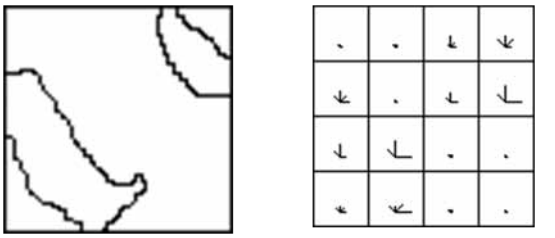


Fig.1 Examples of character form by five writers

A. Character Recognition algorithm[3, 4]

We use the character recognition algorithm "Weighted direction index histogram method [3-5]", which was proposed by our research group, and the method is used widely in many oriental character recognition systems [6].

The feature vector in the recognition method uses the four direction index histogram of the border line of a written character in Fig.2. The size of the feature vector is 64.



(a) border line (b) direction index histograms (feature vector)

Fig.2 Weighted direction index histograms

The discriminate function uses the modified quadratic discriminate function (MQDF). MQDF ${}_l g(x)$ for feature vector x in category l is based on the principal component analysis (PCA), and it uses a mean vector, a set of eigenvectors and eigenvalues per character category as follows.

$$ {}_l g(x) = \sum_{i=1}^{k-1} \frac{\{{}_l \varphi_i{}^t (x - {}_l \mu)\}^2}{{}_l \lambda_i} + \sum_{i=k}^{n} \frac{\{{}_l \varphi_i{}^t (x - {}_l \mu)\}^2}{{}_l \lambda_k} + \ln\left(\prod_{i=1}^{k-1} {}_l \lambda_i \bullet \prod_{i=k}^{n} {}_l \lambda_k\right) \quad (1) $$

where, x: a feature vector of a character sample

${}_l \mu$: mean vector of character in category l

k: the number of used eigenvalues ($k < n$)

n: the size of feature vectors

${}_l \varphi_i$: i-th eigenvector in category l

${}_l \lambda_i$: i-th eigenvalue in category l

t: transpose of a vector

The feature of MQDF is that the number of used eigenvalues is limited as the estimate error of higher eigenvalue has larger error.

The recognition dictionary consisted of the mean vectors, eigenvalues and eigenvectors per category. The mean vector is made from the feature vectors of many characters for learning, and the eigenvalues and eigenvectors are calculated by the covariance matrix of the feature vectors on learning characters.

B. Selection method of a similar writer

We compared between the pure personal dictionary, which is made from the characters written by a specific writer only, and the combined personal dictionary, which is made from the characters written by many writers and a specific writer [1]. The recognition performance of the pure personal dictionary is lower than the combined personal dictionary as the variety of character forms.

We proposed three types of personal character recognition dictionary preciously [1].

(1) "Renewal type dictionary" is renewed repeatedly the mean vector, eigenvalues and eigenvectors by a plenty of characters written by the specific writer and many general writers when the number of written characters increased.

(2) "Modification type dictionary" is renewed repeatedly the mean vectors by the character of the specific writer when the number of written characters increased. However, the eigenvalues and eigenvectors are not renewed, that is, they are the same as the general dictionary made by the characters of many general writers.

(3) "Mixture type dictionary" used the combined the general mean vectors with the mean vector from characters written by a specific writer. The mixed mean vector for a category l and a writer p is given by the following equation.

$$ {}_l \mu_p = \frac{1}{1 + N_p} \left({}_l \mu + \sum_{i=1}^{Np} {}_l f_{p,i}\right) \quad (2) $$

where, N_p is the number of characters written by a specific writer p, ${}_l \mu$ is the feature vector of general dictionary in category l and ${}_l f_{pi}$ is the feature vector of i-th character written by a specific writer p.

The eigenvalues and eigenvectors of "Mixture type dictionary" are the same as the mean vectors of the general dictionary.

The mixture type dictionary was the best solution as the personal dictionary [1]. However, the problem of the mixture type dictionary needs at least one character per category, and the specific writer must write the number of categories. The effort of a user is large as the number of categories is large in the Asian characters.

C. Selection method of a similar writer

In this paper, we propose two new character recognition methods using only one handwritten character (Fig.3). One handwritten character of a specific writer selects the most similar writer among the registered writer. And the selected personal dictionary used to recognize the characters written by the specific writer for all categories. The personal dictionary is made from the set of characters handwritten by the similar writer.

That is the character recognition system calculates the feature vector x of a character by a specific writer. The system recognizes the feature vector by the specific writer using all personal dictionaries (registered writer number: n=1, 2, … , N) in the same written category l (Fig.4), and it selects the most similar writer using the least value of MQDF ${}_lg(x, n)$. The system uses the selected personal dictionary as the dictionary of the specific writer for all categories.

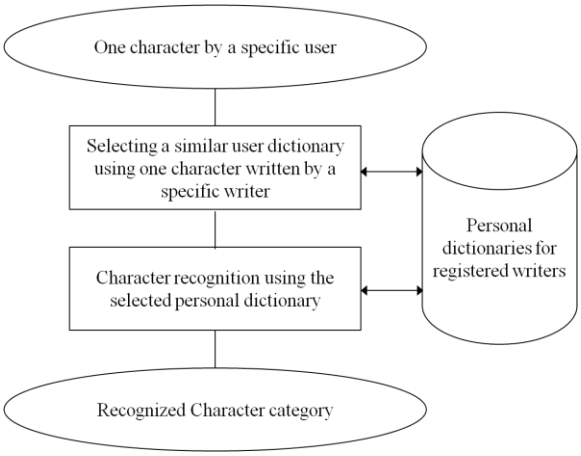


Fig.3 The general flowchart of a proposed recognition process

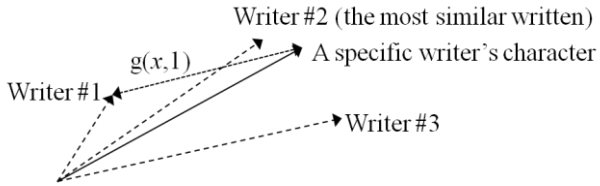


Fig.4 Selecting the most similar writer in feature space

D. Similar mean dictionary(SMD) and Similar feature space dictionary(SFSD)

In this paper, we propose two new personal recognition dictionaries, "Similar mean dictionary (SMD)" and "Similar feature space dictionary (SFSD)".

SMD consists of the combined mean vector of characters written by the general writer and the similar writer (similar mean vector), and the set of the eigenvalues and the eigenvectors of characters written by the general writer. The similar mean vector ${}_l\mu_{sm}$ is calculated by the combination of characters written by the general writer and the similar writer, and the generating procedure of the similar mean vector is as follows.

$$ {}_l\mu_s = \frac{1}{Ns}\sum_{i=1}^{Ns} {}_l f_{s,i} \tag{3} $$

$$ {}_l\mu_{sm} = \frac{Ns}{1+Ns}\,{}_l\mu_s + \frac{1}{1+Ns}\,{}_l\mu \tag{4} $$

where, Ns : the number of characters written by a similar writer

${}_l\mu$: the feature cector of general dictionary in category l

The similar mean vector is approaching to the personal mean vector as the number of learning characters written by the similar writer is increasing.

SFSD consists of the similar mean vector of characters written by the general writer and the similar writer, and also the set of the eigenvalues and the eigenvectors of characters written by the general writers and the similar writers.

The mean vector of the SFSD is the same as SMD. However, the set of eigenvalues and eigenvectors is not the same with the SMD, and it is calculated by the covariance matrix on the feature vectors of characters written by the general writers and the similar writers.

Two new proposed dictionaries use one character written by a specific writer, and the effort of the writer is the minimum to reflect the handwritten feature of a specific writer. The comparison of four dictionaries is showed in Table 1

Table 1 The comparison of four dictionaries

Type of recognition dictionary	The number of characters written by a specific writer	The writers of mean vector	The writers of eigenvlues and eigenvectors
General	0	general	general
Mixture type	71	general+ specific	general
Similar mean	1	general+ similar	general
Similar feature space	1	general+ similar	general+ similar

III. Experimental results

We compared the effect of these four dictionaries. A set of test samples is handwritten Japanese "HIRAGANA"

characters by ten writers. We prepared the ten characters for each "HIRAGANA" character category by pen tablet "Cintiq Interactive Pen Display" by Wacom Co. LTD. The resolution of the character was about 100 x 100 pixels. We didn't use the time information and used the image only. As the feature vector we used the histograms of 4x4meshes, 4 directions, that is, 64 dimensional feature vector.

The general dictionary was made from the "Hiragana" characters written by 200 writers per one category in the Japanese character data base ETL9B. The database was established by the Electro-Technical Laboratory (ETL) of Japan [at present, the National Institute of Advanced Industrial Science and Technology (AIST) of Japan].

Three personal dictionaries (mixture type dictionary, similar mean dictionary (SMD) and similar feature space dictionary (SFSD)) were made from ten characters per category written by ten writers and ETL9B. One handwritten character written by a specific written is a "pa" in "HIRAGANA".

We experiment using *"leave one out method"* for writers to obtain the recognition rates, that is, if the writer "a" would be a specific writer, then, the most similar writer is selected among the other writers. The experiment was repeated for every writer.

Fig.5 shows the comparison of four dictionaries in correct recognition rates for ten writers. The rates of mixture type dictionary and SFSD are nearly equal, and these rates are better clearly than the general dictionary for all writer. The rate of SFSD is better than the rate of SMD. The rate of SMD is better than the general dictionary.

SFSD obtained the best recognition rate 91.0% relatively to the rate of ordinal general dictionary 82.4% in Table 2.

We confirmed that only one character by a specific writer is very effective for handwritten character recognition.

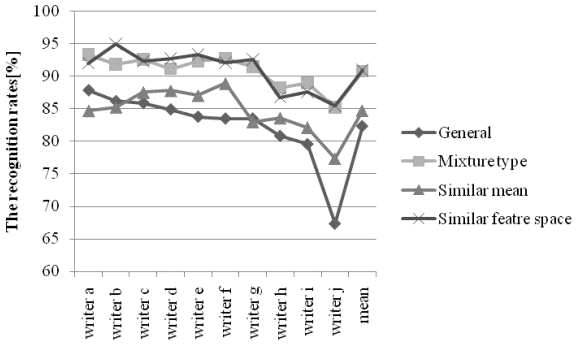


Fig.5 The recognition rates by four dictionaries

Table 2 The comparison of correct recognition rates

Type of recognition dictionary	correct recognition rate[%]
General	82.4
Mixture type	90.8
Similar mean	84.7
Similar feature space	91.0

IV. Conclusions

We proposed two new generating methods: "similar mean dictionary (SMD)" and "similar feature space dictionary (SFSD)" to improve the personal character recognition. These dictionaries use only one character, and they can estimate the personality of writing style using characters written by a similar writer. We examined handwritten characters including 71 kinds of Japanese "HIRAGANA" characters for each writer, and we obtained the character recognition rate of 91 % (the general dictionary made from ETL9B: 82 %).

We confirmed that only one character by a specific writer is very useful for handwritten character recognition.

In the future, selection method of a similar writer changes to select several similar writers from one person, and synthesizes several feature vectors. We improve the accuracy of character recognition by using the synthesized feature vector.

References

[1] S. Tsuruoka, H. Morita, F. Kimura, and Y. Miyake, Handwritten Character Recognition Adaptable to the Writer, *Trans. IEICE*, Vol.J70-D, pp.1953-1960, October 1987 [in Japan]

[2] S. Thuruoka, H. Morita, F. Kimura and Y. Miyake, Handwritten Character Recognition Adaptable to the Writer, *Proc. of IARP Workshop on Comoputer Vision*, pp.179-182 October 1988

[3] S. Tsuruoka, K. Kurita, T. Harada, F. Kimura, and Y. Miyake, Handwritten "KANJI" and "HIRAGANA" Character Recognition Using Weighted Direction Index Histogram Method", *Trans. IEICE*, J70-D, pp.1390-1397, July 1987 [in Japanese]

[4] F. Kimura, K. Takashina, S. Tsuruoka, and Y. Miyake, Modified Quadratic Discriminant Functions and the Application to Chinese Character Recognition, *IEEE Trans. PAMI-9*, pp.149-153, January 1987

[5] F. Kimura, T. Wakabayashi, S. Tsuruoka, and Y. Miyake, Improvement of Handwritten Japanese Character Recognition Using Weighted direction Code Histogram, *Pattern Recognition*, vol. 30, no.8, pp.1329-1337, Aug. 1997

[6] M. Cheriet, N. Kharma, C. Liu, C. Y. Suen, Character recognition systems –A Guide for Students and Practitioners-, pp.59-63, pp. 134 – 135, pp. 292- 301, *Wiley & Sons, Inc.* , 2007