IAPR Workshop on CV - Special Hardware and Industrial Applications OCT.12-14, 1988, Tokyo

HANDWRITTEN CHARACTER RECOGNITION ADAPTABLE TO THE WRITER

Shinji TSURUOKA+, Hiroyuki MORITA+*,
Fumitaka KIMURA+, and Yasuji MIYAKE+

+Faculty of Engineering,Mie University,1515 Kamihama-cho,Tsu-shi,Mie 514,Japan
*NIPPONDENSO CO.,LTD.,Kariya-shi,Aichi 448 Japan

ABSTRACT

In this paper, we describe **the handwritten character recognition adaptable to the writer.** It is efficient when the specific writer uses the same OCR for many characters. At the early stage, input characters are recognized using general dictionary, and then the correctly recognized character modify the dictionary to be adaptable to the variation of the characters of the specific writer. Using the adaptable dictionary modified by 10 characters/category, the classification rate is improved from 96.8% (general dictionary) to 99.5%.

1. INTRODUCTION

The process of the automatic recognition of documents and line drawings consists of three subjects. The first is separation between the part of lines and figures and the part of characters. The second is recognition of line attribute (type, length, angle,connectivity and so on) and recognition of figures. The third is character recognition .

This paper discusses the third subject. One document or line drawing is ordinarily written by one person. So, the written character font is similar to each othor. In Optical Character Reader for personal use (**Personal OCR**), this fact enables to achieve higher recognition rate. It learns only single user's characters.

In this paper, we propose the handwritted character recognition adaptable to the writer. It is the efficient, when the specific writer uses the same OCR for many characters. At the early stages, input characters are recognized using general dictionary, which is designed by many generic writers' characters, and then the input characters of specific writer modify the general dictionary to be adaptable to the writer.

By using this method,personal OCR can correctly recognize the characters peculiar to writer, for example, the character with contacted strokes or lengthy strokes.

This paper proposes 3 types of adaptive character recognition (renewal type, modification type, mixture type) which use mean vector,eigenvalues and eigenvectors. Using "Weighted Direction Index Histogram Method (WDIHM)[1],[2]", we made the OCR system recognize the sets of "HIRAGANA" characters (46 categories) written by spesific writer, and found that using the adaptable dictionaries which were designed by 10 characters/category, the classification rate was improved from 96.8% (general:designed by 100 characters/category in data base ETL4) to 99.5%. These experiments proved these methods to be effective.

The result of the comparison among 3 types of adaptive dictionaries shows that mixture type dictionary (mean vector is composed of general mean and personal mean, and eigenvalues and eigenvectors are the same as general dictionary) gives higher recognition rate and requires less calculation time and storage (800 k Byte/person) for modifying the dictionary.

2. PERSONAL DICTIONARY AND GENERAL DICTIONARY

It is known that the recognition rate is improved using a personal dictionary, which is designed by writer's characters only,in comparison with using general dictionary[3],[4]. We investigate this property in " Weighted Direction Index Histogram Method (WDIHM) ". This OCR system obtains the classification rate of 99.5% (high quality sets) and 96.3% (low quality sets) in 927 categories (Data base ETL8)[1], which is the top level classification rate at present .

2.1 FEATURE EXTRACTION

We employ the "WDIHM" as feature extraction. The procedure consists of (1)Binarization,(2)Normalization of position and size, (3)Border following and 4-direction index coding,(4)Generation of histograms of index code in each 7(vertically)*7(horizontally) subregions, and (5)Compression of 7*7*4 histograms into 4*4*4 histograms (64 dimension) with gaussian filter(Fig.1).

2.2 DISCRIMINANT FUNCTION

We use "Modified Quadratic Discriminant Function(MQD

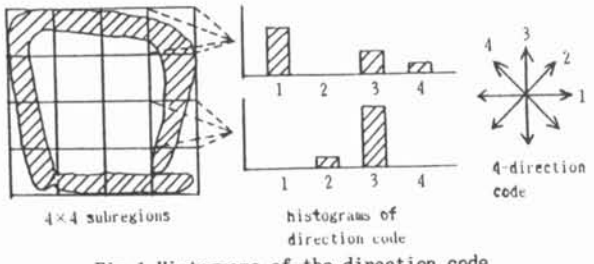

Fig.1 Histograms of the direction code

F)"[1]. The MQDF for category l is given as

$$g^{l}(x)=\sum_{i=1}^{k}\frac{(x-{}^{l}\mu,\ {}^{l}\Phi_i)^2}{{}^{l}\lambda_i}+\sum_{i=k+1}^{n}\frac{(x-{}^{l}\mu,\ {}^{l}\Phi_i)^2}{{}^{l}\lambda_{k+1}}+\ln\left(\prod_{i=1}^{k}{}^{l}\lambda_i\cdot\prod_{i=k+1}^{n}{}^{l}\lambda_{k+1}\right)$$

$$=\frac{1}{{}^{l}\lambda_{k+1}}\left\{\|x-{}^{l}\mu\|^2-\sum_{i=1}^{k}\left(1-\frac{{}^{l}\lambda_{k+1}}{{}^{l}\lambda_i}\right)(x-{}^{l}\mu,\ {}^{l}\Phi_i)^2+\ln\left(\prod_{i=1}^{k}{}^{l}\lambda_i\cdot\prod_{i=k+1}^{n}{}^{l}\lambda_{k+1}\right)\right\}. \qquad (1)$$

where, x : feature vector of input character
$^{l}\mu$: mean feature vector of category l
$^{l}\lambda_i$: ith eigenvalue of covariance matrix $^{l}\Sigma$
$^{l}\Phi_i$: ith eigenvector of covariance matrix $^{l}\Sigma$
k : integer ($1\leq k\leq m,n$)

The properties of this discriminant function are that (1)classification rate is higher (2)calculation time, storage and calculation error are less than a quadratic discriminant function.

2.3 EXPERIMENT

(1)SAMPLE : Specific writer's characters are 26 characters/category·writer written with 0.5mm mechanical pencil on OCR sheets by 5 undergraduates . Character size is about 40*50 dots. Generic writer's characters for general dictionary are characters in data base ETL4, which is made in Electrotechnical Laboratory. This is 100 characters/category and 1 character/writer written similarly.

Personal dictionaries are made by 16 charcters/category and the rest (10 characters/category) is used for test sample. General dictionary is madeby 100 characters/category.

(2)CLASSIFICATION RATE : The classification rates of 5 writers are shown in Fig.2. General dictionary have wide variation and the mean classification rate is 96.8%, that is less than personal dictionary (99.0%). The misrecognized characters are gathered in some specific categories. This fact is shown that specific writer's conceptual patterns in some categories are different from the means of other writers' conceptual patterns.

(3)MISRECOGNIZED CHARACTER :Examples of misrecognition with personal dictionary, which are correctly recognized with general dictionary, are shown in Fig.3. It shows that the characters which are written by writer in the unusual font are misrecognized. This fact shows that the personal dictionary is adaptable only to the little variation of character font, because it is designed by the specific writer's characters only.

3. HANDWRITTEN CHARACTER RECOGNITION ADAPTABLE TO THE WRITER

"Handwritten character recognition adaptable to the writer" means that the OCR system uses general

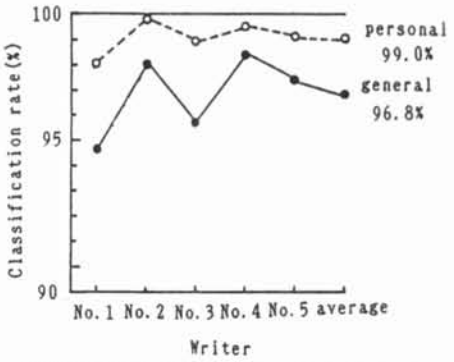

Fig.2 Comparison between personal dictionary and general dictionary

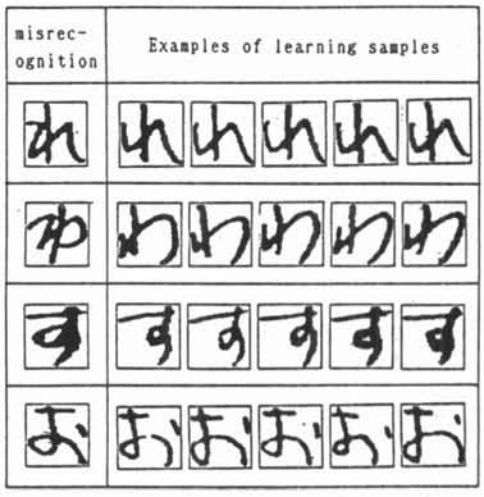

Fig.3 Examples of misrecognition by personal dictionary

dictionary at the early stages, and that dictionary is adaptable to the specific writer according as it learns the writer's font , so that it uses personal dictionary at the stages of enough learning.

3.1 TWO TYPES OF ADAPTABLE DICTIONARY

3.1.1 TYPES OF DICTIONARY

(1)RENEWAL TYPE DICTIONARY : Renewing the dictionary (mean vector, eigenvalues, and eigenvectors) from the characters for making general dictionary and the characters of specific writer.

(2)MODIFICATION TYPE DICTIONARY : mean vector is renewed from the specific writer's characters,and eigenvalues and eigenvectors leave it alone,namely they are eigenvalues and eigenvectors of general dictionary.

Renewal type is natural in statistics, and it is standard dictionary. But it must store the mean vector of generic writers' and specific writer's characters, and the covariance matrix of feature vectors for eigenvalues and eigenvectors.

Modification type is unnatural as mean vector only is renewed. But renewing eigenvalues and eigenvectors is practically impossible because the sample size is too small. The covariance matrix in general dictionary is reliable in comparison with them in personal dictionary.

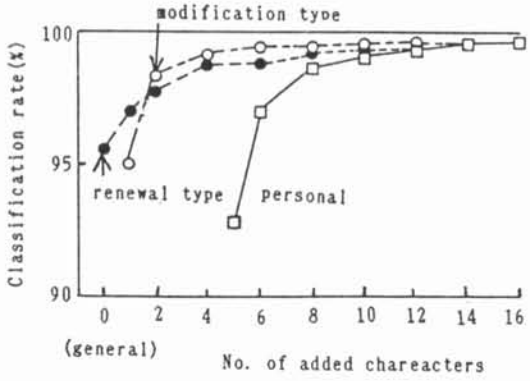

Fig.4 Comparison between renewal type and modification type

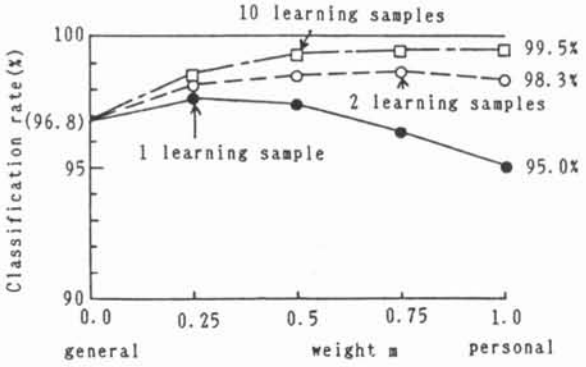

Fig.5 Classification rate vs. weight

3.1.2 RESULTS

(1)RENEWAL TYPE : The relation of classification rate vs. number of personal characters in renewal type dictionary is shown in Fig.4. Classification rate in case of 0 character means the rate in general dictionary (96.8%), and this result prove that the classification rate improves according to adding personal characters.

Mean classification rate of 5 writers in renewal type is 99.3% in 10 characters learning , it is higher than 99.0% in personal dictionary. This property shows that renewal dictionary have various deformation in the category, as it includes personal characters and many writers' characters . We think that this property is independent of feature extraction because it hold in pattern matching method[5].

(2)MODIFICATION TYPE : The relation of classification rate vs. number of personal characters in modification type dictionary is shown in Fig.4. Classification rate in case of 1 character is lower than general dictionary, and there are 100% error in some categories . This shows that one character of the writer can't represent a standard font in a category.

(3)COMPARISON : Classification rates of renewal and modification type are higher than that of personal dictionary. In case of 1 character, the rate of renewal type is higher than that of modification type, but in case of more characters, modification type is higher than renewal type. Saturation of these rise at about 8 characters.

3.2 MIXED ADAPTABLE DICTIONARY

We wish that actual OCR system obtains high classification rate in few samples. In this point, we think that modification type is desirable except case of 1 character . So that, we propose the new method which makes mean vector in modification type considering mean vector in general dictionary.

3.2.1 RECOGNITION BY THE WEIGHTED MEAN VECTORS

The weighted mean vector ${}_w\overline{f}^l$ can be defined by Eq. (2) which is made from specific writer's mean vector ${}_p\overline{f}^l$ and mean vector ${}_g\overline{f}^l$ in general dictionary.

$$ {}_w\overline{f}^l = (1-m)\,{}_g\overline{f}^l + m\,{}_p\overline{f}^l \qquad (2) $$

where, m : weight ($0 \leqq m \leqq 1$)

The weighted mean vector makes mean vector in general dictionary when m=0 , and it makes mean vector in modification dictionary when m=1.

The relation of classification rate and weight m is shown in Fig.5 . As a result of this experiment, we wish that weight is not constant but decrease according to increase of the learning samples.S.Naito et al.[4] showed similar result in the storoke density function method, but it dosen't use eigenvalues and eigenvectors,so in few learning samples classification rate falls abruptly when the weight of personal vector is large.

3.2.2 MIXTURE TYPE DICTIONARY

We propose " **mixture type dictionary** ", which uses blend mean vector ${}_b\overline{f}^l$ in Eq.(3) , eigenvectors and eigenvalues in general dictionary.

$$ {}_b\overline{f}^l = \frac{1}{{}_pN+1}\left({}_g\overline{f}^l + \sum_{i=1}^{{}_pN} {}_pf^l{}_i\right) \qquad (3) $$

where, ${}_g\overline{f}^l$: mean vector of category l in general dictionary

${}_pf^l{}_i$: ith feature vector of category l in writer(p)'s learning samples

${}_pN$: Number of writer's learning samples

Eq.(3) means that mean vector in general dictionary is treated as one writer's learning sample. Varying ${}_pN$ as 1,2,10 corresponds to varying weight m as 0.5, 0.7, 0.9 , and this method is one of the realizable method about the preceding result.

Classification rates in this method are shown in Fig.6. In all writers exept No.1, classification rate in only one learning character is higher than general dictionary , so that it is proved that this method is effective in OCR.

Comparision of dictionary types is shown in Fig.7. Classification rates in mixture type dictionary are higher than others, and we think that the defect of modification type is improved.

3.3 COMPARISON OF 3 TYPE DICTIONARIES

(1)RECOGNITION RATE : Classification rate in renewal type is higher than that in general dictionary. Classification rate in modification type is lower than

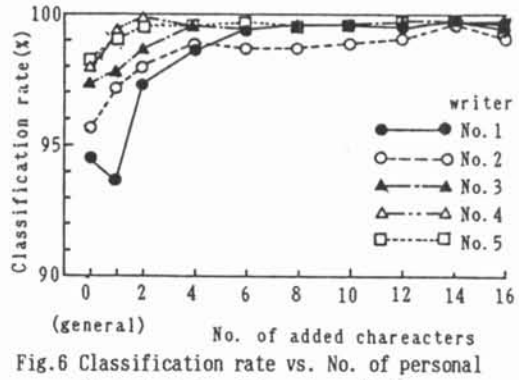

Fig.6 Classification rate vs. No. of personal characters in mixture type dictionary

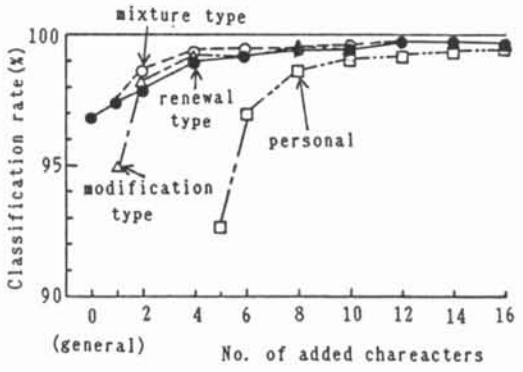

Fig.7 Comparison of dictionary types

that in general dictionary in case of 1 learning character, but in other cases the rate is higher than that of renewal type. In mixture type, classification rate is highest among them in every case.

(2)CALCULATION TIME : Calculation time of Renewal type is very great as system calculates mean vector, covariance matrix, eigenvalues and eigenvectors at every learning character. As compared with it, modification and mixture type calculate mean vector only, so they have small calculation. To reduce the calculate time, mean vectorcan be calculated by the recursion relation [6] in a little deformation. The actual calculation time are 140.7 ms/character in renewal type, 0.1 ms/character in modification and mixture type by FACOM M 382 in Nagoya University Computation Center. Most of the time is the calculation of eigenvalues and eigenvectors.

(3)STORAGE SIZE : In modification and mixture type, it hold mean vectors only. But in renewal type it must hold mean vector and covariance matrix.The ratio is 65 in 64-dimensional feature vector. In case of using 4 Bytes real number, storage size in modification and mixture type is 256 Bytes/category. At 3,000 category (JIS 1-st level Chinese Characters) it is 800 k Bytes, and 1 M Bytes floppy disk can include it. Writer can use specific personal dictionary in floppy disk when he uses the OCR system.

As a result of these discussions, we think that mixture type dictionary is most effective among 3-types.

4. CONCLUSION

In this paper, we proposed three types of dictionaries (renewal, modification, mixture) adaptable to writer. The followings were shown by experiments.

(1)For personal dictionary, the classification rate (99.0%) is higher than that (96.8%) in general dictionary.

(2)For adaptable dictionary, the classification rate (renewal:99.3%, modification:99.5%, mixture:99.5% in 10 characters learning) is higher than that (99.0%) of personal dictionary when number of specific writer's learning samples are same.

(3)Mixture type dictionary has higher calassification rate, less calculation time and storage size (800 k Bytes). We think that this type of adaptable dictionary is most practical in OCR system.

The followings remain to be studied in the future.

(1)Applying this method to Chinese character , and printted or dotprintted character recognition.

(2)Extracting personality independent of category from characters of partial categories , and using the personality effectively to recognize characters in unknown categories[7].

ACKNOWLEDGMENT

We would like to sincerely thank Prof. M.Yoshimura for her helpful coments and offering personal characters, and members of Image Processing Section in Electrotecnical Laboratory for use of image scanner and data base ETL4.

REFERRENCES

(1)S.Tsuruoka,M.Kurita,T.Harada,F.Kimura,and Y.Miyake: "Handwritten "KANJI" and "HIRAGANA" Character Recognition Using Weighted Direction Index Histogram Method",Trans.IEICE Japan,vol.J70-D,7,pp.1390-1397, July 1987.

(2)F.Kimura,K.Takashina,S.Tsuruoka,and Y.Miyake:"Modified Quadratic Discriminant Functions and the Application to Chinese Character Recognition" , IEEE Trans.Pattern Anal.Machine Intell.,vol.PAMI-9,1,pp. 149-153,January 1987.

(3)M.Yoshimura,F.Kimura,I.Yoshimura:"On the Effectiveness of Personal Templates in the Character Recognition ", Trans.IECE Japan, vol.J66-D,4,pp.454-455, April 1983.

(4)S.Naito,and I.Masuda:"Chinese Character Recognition Based on Personal Handwriting Characteristics " , Trans.IECE Japan,vol.J67-D,4,pp.480-487,April 1984.

(5)H.Morita,S.Tsuruoka,F.Kimura, and Y.Miyake :" Handprinted Character Recognition Fitted to the Writer (I) ",IECE Tec.Rep.Japan, PRL84-23,July 1984.

(6)R.O.Duda and P.E.Hart: "Pattern Classification and Scene Analysis" , pp.82, A WILEY-INTERSCIENCE PUB., 1973.

(7)S.Tsuruoka,O.Taniguti,F.Kimura,and Y.Miyake: "Handwritten Character Recognition Using Individual Vector",Nat.Conf.Rec.I&S IEICE Japan,pp.65, Nov. 1987.