Large Scale Hand-written Character Recognition System Using Subspace Method

Nei KATO and Yoshiaki NEMOTO
Graduate School of Information Sciences
Tohoku University
Aobayama, Aoba-ku, Sendai 980-77, Japan

Abstract In spite of the fact that subspace method can approximate the distribution of categories precisely, only a few attempts have so far been made at applying it in hand-written character recognition. The subspace method proposed by Watanabe offers the basic concept of subspace construction, but the issue of how to use the limited samples to construct effective subspace to avoid the problem of the subspace inclining toward mean vectors remains unresolved. To cope with this problem, the authors have proposed the Combination method(CM)[1], which constructs the subspace from several groups including different number of samples devided from the whole training samples. CM obtained a high recognition rate of 97.76% with respect to ETL9B, the largest database of hand-written characters in Japan; the issues that need to be dealt with next are how to improve the recognition accuracy and how to accelerate the recognition speed. In this paper, we propose a new method called Uniform Division Method(UDM), which uses the uniformly divided training samples to construct subspace. Compared to CM given earlier, UDM is very simple and effective enough to improve the accuracy of recognition; as a result, we obtained a recognition rate of 98.64% for ETL9B compared to the 97.76% for CM. This is the first time that such a high recognition rate has been obtained by making good use of subsapce method. Furthermore, the computation required for UDM is less than a half of that of CM. The UDM algorithm and the experiments with ETL9B will be described in this paper.

1 Introduction

Subspace method is well-known for its capability to approximate the distribution of categories precisely[2]. Due to the limitation of computation resources, so far the application of subspace method to large scale hand-written character recognition has been superficial. In recent years, with the advance of computer hardware, several articles on distinguishing between similar Chinese characters[3] and segmentation[4] have used the subspace method.

The subspace method proposed by Watanabe offers the basic concept of subspace construction, but the issue of how to use the limited samples to construct effective subspace to avoid the problem of the subspace inclining toward mean vectors[5, 6] remains unresolved. To cope with this problem, the authors have proposed the Combination method(CM), which constructs the subspace from several groups including different number of samples divided from the whole training samples. This idea is based on the multi-template concept[7, 8] for hand-written character recognition, in other words, the technique employed involves preparation of multi-subspaces for each character. Although CM obtained a high recognition rate of 97.76% with respect to ETL9B, the largest database of hand-written character in Japan, the problem of repeated use of samples remains. This leads to the deterioration of recognition ability as well as the recognition speed. Obviously, the question which we must consider next is to improve the recognition accuracy and to accelerate the recognition speed.

In this paper, we propose a new method called Uniform Division Method(UDM), which uses uniformly divided training samples for subspace construction. First, we verify that UDM based subspace method possesses adequate ability, in comparison with other well-known clustering methods such as K-means method or the modified K-means method in which the number of character is fixed. Then we show that the features of characters can be represented by fewer orthonormal vectors when the training sample is divided into several groups. This method of division improves on the drawback encountered in CM by a large margin. In experiments

0-7803-3280-6/96/$5.00 ©1996 IEEE

of ETL9B, we obtained a very high recognition rate of 98.64%, in comparison with the 97.76% of CM. Furthermore, the computation was reduced to less than half of that in CM. ¿From these results, the proposed method can be considered very effective in handwritten character recognition.

The paper is organised as follows. Section 2 gives a brief overview of the conventional subspace method, and then introduces the new UDM, and discusses the results of basic simulation and experiments. Section 3 describes the handwritten character recognition system. Section 4 presents the evaluation results of experiments on the ETL9B and related discussion. This is followed by concluding remarks in Section 5.

2 Uniform Division of Training Samples for Subspace Method

In this section, firstly the problem of conventional subspace method will be pointed out, and then the three approaches to cope with the problem will be described. Subsequently, we present UDM and confirm its effectiveness by simulation and small scale recognition experiments.

2.1 Problems of Conventional Subspace Method

To see the problem of conventional method, we conduct the following experiment. One set of 200 samples pertaining to one HIRAGANA じ were selected from the ETL9B Database. In the experiment, the 200 samples are divided into 10 groups, each of them has 20 sets data. Leave-one-out[9] method is employed here, which means 9 groups are used for subspace construction and the remaining one is used as unknown data.

Fig. 1 shows the recognition result for HIRAGANA じ, it uses the 10 eigenvectors which have the largest eigenvalue.

Diamond "◇" and plus "+" designate correct and wrong respectively. The horizontal axis represents sample number, and the vertical axis represents the distance transformed from projection of input vectors on subspace in terms of the Euclidean vector norm. Let i denote the sample number and P_i^m its projection on subspace, P_{max}^m the maximum projection of categories m; transformed distance d_i^m is defined as

$$d_i^m = P_{max}^m - P_i^m \tag{1}$$

where i belongs to categories m.

We can see from Fig.1 that the performance of conventional subspace method is not good for samples far from mean-vectors. In other words, in the conventional method of subspace construction samples far from the mean-vectors are not reflected efficiently into subspace.

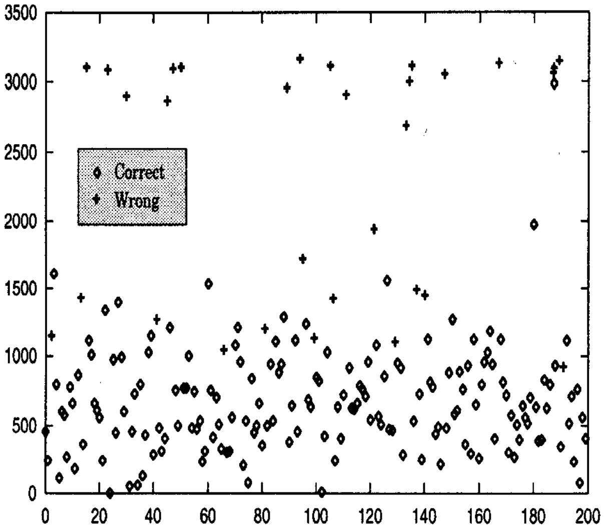

Figure 1: Distance distribution of recognition results of HIRAGANA じ

2.2 Basic Experiments in Division Method of Subspace

In section 2.1, we showed that subspaces constructed by conventional method incline toward mean-vectors, and that wrong classification of samples far from the mean-vectors occur often. This is the main problem that should be overcome. To cope with this problem, the authors' idea is to divide the training samples into several groups; this strongly influences recognition of the samples faraway in the subspace.

There are three possible methods of division. Here, we are going to examine the effect of each method through experiments. The test case is the one where the training samples are divided into two.

Method 1: Uniform Division Method(UDM)

Method 2: Modified K-means method with fixed number in each group

Method 3: K-means method[10]

Method 1 simply divides the training samples into halves from the beginning sample.

Method 2 first divides the training samples as Method 1, then uses K-means method to re-allocate sample data so that within-cluster variance becomes smallest.

Method 3 is K-means method with the limitation that sample number in each group must be larger than

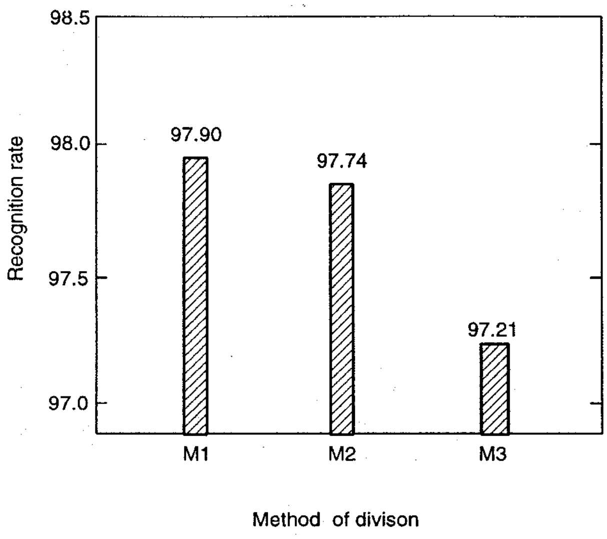

Figure 2: Results by using 3 different division methods

40. The reason is that the subspace can not be constructed effectively with numbers below 40.

Figure 2 shows the recognition results of the 3 division methods. M1, M2, M3 indicate Method 1, Method 2, Method 3 respectively. In this paper, unknown samples were 20 sets selected as 1, 11, 21, $\cdots$, 191. The remaining data was used as training samples. We used the following classification rule for UDM:

if for all $j \neq i$

$$\sum_{m=1}^{2} x^t P_m^{(i)} x > \sum_{m=1}^{2} x^t P_m^{(j)} x \qquad (2)$$

then classify x in category i.

In eq.(2), m denotes subspace number of category i. $P_m^{(i)}$ denotes the $m - th$ projection matrix of category i.

$$P_m^{(i)} = \sum_{k=1}^{p^{(i)}} u_{m,k}^{(i)} u_{m,k}^{(i)T} \qquad (3)$$

where $u_{m,k}^{(i)}$ are orthonormal vectors. The eigenvectors are chosen to correspond to the largest eigenvalues for $p^{(i)}$. $p^{(i)}$ is subspace dimension. The value of $p^{(i)}$ was chosen as 40 for M1, 38 for M2 and 35 for M3 with best recognition accuracy.

We can see from Fig.2 that there is little difference between Method 1 and Method 2. On the other hand, Method 3 shows relatively low performance. The drawback of k-means method is that some groups contain only a few characters and that leads to an imbalance between subspaces.

To see the effectiveness of UDM in more detail, simulation and recognition experiments were conducted. We will discuss these in the next section.

2.3 Proper Division Number for UDM

In this section, we discuss the proper division number for UDM. To see the relation between the division number and recognition rate, the following experiment is conducted. As in section 2.2, the 20 unknown sample sets are selected as 1, 11, 21, $\cdots$, 191 from the 200 ETL9B sets. The remained 180 sets are used as training sample sets. The results are shown in Fig.3. The horizontal axis denotes the number of eigenvectors which correspond to the largest eigenvalues. The vertical axis denotes the recognition rate. The 3 line graphs indicated by square, dot and triangle denote division number 3, 4 and 5 respectively. From Fig.3, we can see that the division number and eigenvectors are closely related since the highest points of recognition rate keep changing when division number varies. To sum up, when division number increases, the samples included in one group decrease, this leads to the concentration of features to eigenvectors with larger number. This is the reason that point of highest recognition rate moves left when number of division increases. From Fig.3, the division number 4 and eigenvector number 25 can be considered the best combination for UDM.

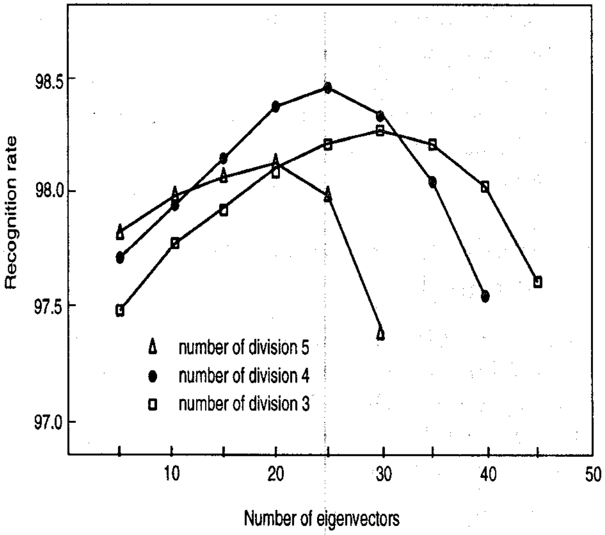

Figure 3: Relations between number of division and eigenvectors

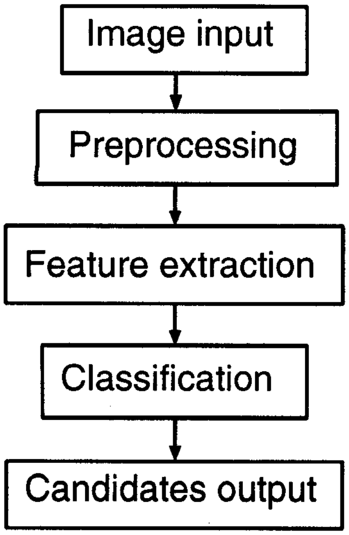

Figure 4: Handwritten character recognition system

3 Scheme of Hand-written Character Recognition

The flowchart of recognition system is given in Fig.4. The system consists of three procedures, namely, preprocessing, feature extraction and classification.

3.1 Preprocessing

The preprocessing consists of smoothing, noise reduction and normalization[11].

3.2 Extraction of Feature Vectors

Improved Directional Element Feature[1] is used as characteristic feature. It derives from Directional Element Feature[12] which is originally the characteristic feature for printed character.

3.3 Classification

Going by the recognition speed, the classification is further divided into two procedures, rough classification and fine classification. In rough classification, city-block distance is used as distance measure, and the first 30 candidates are selected. In fine classification, we use the proposed UDM on subspace method.

4 Evaluation by Using Database ETL9B

4.1 Contents of Experiments

In this section, the experiment using the whole ETL9B data sets will be described. In the experiment, the training sample sets and unknown sets are assigned as in Table 1. The 10 groups are named from A to J. Division number of training samples and number of eigenvectors are selected as 4 and 25 respectively in accordance with the results from experiment detailed in section 2.3.

Table 1: Division of ETL9B sets

Groups	Unknown sets	Training sets
A	1 ~ 20	Remained 180 sets
B	21 ~ 40	Remained 180 sets
C	41 ~ 60	Remained 180 sets
D	61 ~ 80	Remained 180 sets
E	81 ~ 100	Remained 180 sets
F	101 ~ 120	Remained 180 sets
G	121 ~ 140	Remained 180 sets
H	141 ~ 160	Remained 180 sets
I	161 ~ 180	Remained 180 sets
J	181 ~ 200	Remained 180 sets

Table 2: Recognition rate of ETL9B

Groups	1st	2nd	3rd
A	99.06	99.71	99.81
B	98.58	99.47	99.62
C	98.94	99.65	99.76
D	98.65	99.58	99.73
E	98.66	99.56	99.72
F	98.50	99.45	99.60
G	98.40	99.36	99.53
H	98.59	99.58	99.73
I	98.52	99.50	99.65
J	98.47	99.56	99.69
Ave.	**98.64**	**99.54**	**99.68**

4.2 Results of Experiments

The average recognition rates of each groups are shown in Table 2, and the recognition rate of each set is shown in Fig.5 for reference. From Table 2, we can see that a

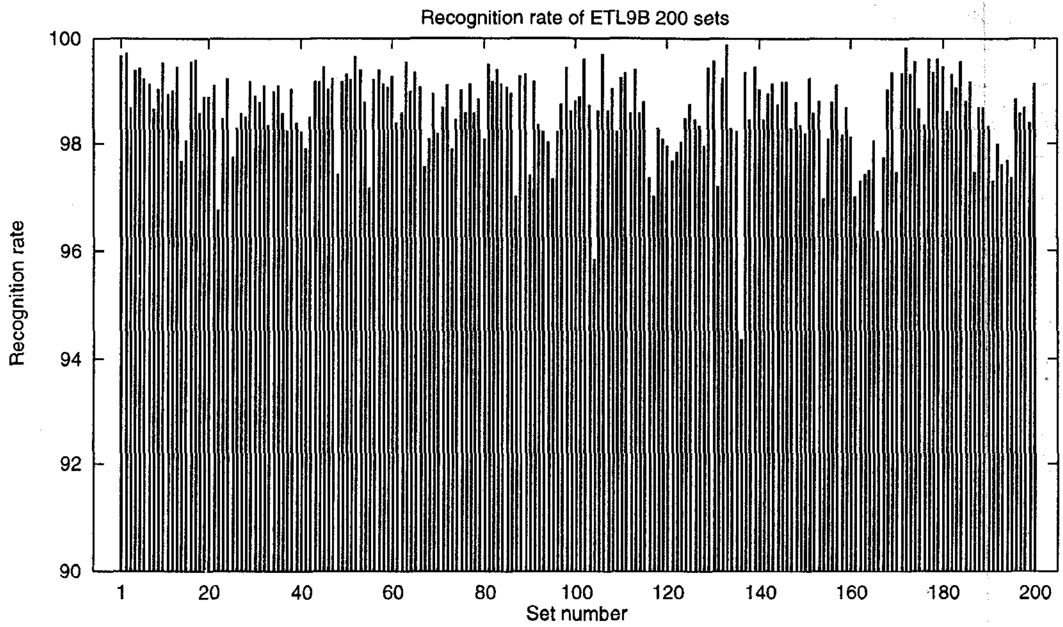

Figure 5: Recognition rate of ETL9B sets

very high recognition rate of 98.64% was obtained by the proposed UDM.

4.3 Computation Time

In comparison with CM proposed in [1], the number of multiplying eigenvectors by unknown vector is reduced from 240(6 subspaces × 40 eigenvectors) to 100(4 subspaces × 25 eigenvectors), thereby the recognition speed is more than doubled.

5 Conclusion

The conventional subspace method offers the basic concept of subspace construction, but the issue of how to use the limited samples to construct effective subspace to avoid the problem of the subspace inclining toward mean vectors remains unresolved. In this paper, to cope with this problem, we proposed Uniform Division Method(UDM), which basically divides the training samples into several groups having equal numbers. We have proved that using the UDM can remarkably improve the recognition accuracy as well as the recognition speed. In the experiment with ETL9B, the proposed method gave a very high recognition rate of 98.64%. This recognition rate is superior to the Combination method(97.76%), which was proposed by the authors in [1]. The recognition speed also more than doubled compared to that of CM. All these results confirm that the proposed UDM is very effective in hand written character recognition.

Future research efforts include recognition techniques for similar KANJI characters as well as further improvement of recognition speed.

Acknowledgment

The authors would like to thank NEC Corporation for providing us the EWS4800/360 workstation and Electro technical Laboratory for providing us the ETL9B Database.

References

[1] Abe, M., Sun, N. and Nemoto Y.(1995), A Hand-written Character Recognition System by Using Improved Directional Element Feature and Subspace Method, IEICE Trans. Syst. & Info., Vol.J78-D–II, No.6, pp.922-930.

[2] Watanabe, S.(1969), Knowing and Guessing, John Wiley & Sons.

[3] Kotani, H., Hasimoto, R., Ohkura, M. and Shiono, M.(1992), Similar Hand-written Character Recognition by using Multi-subspace Method, Proc. of Meeting on Image Recognition and Understanding, vol.1, pp.215-222.

[4] Motegi, Y. and Ariki, Y., "Segmentation of Hand Written Character Using Subspace Method," *IEICE Technical Report*, PRU94-97, 1995.

[5] E., Oja, *Subspace Methods of Pattern Recognition,* Research Studies Press, England, 1983.

[6] Mori, K., *Pattern Recognition*, The Institute of Electronics, Information and Communication, 1988.

[7] Hai, T., Morishita, T., Kabuyama, Y., Izaki, Y. and Yamamoto, E., "A Study of Multiple Template Matching for Handwritten KANJI Character Recognition," *IEICE Technical Report*, PRL81-42, 1981.

[8] Shiono, M., "Basic Experiments of Handwritten Character Recognition by Using Multi-dictionary Similarity Method," *Trans. of IPSJ*, vol.27, no.9, pp.853-859, Sept. 1986.

[9] Lachenbruch, P.A. and Michey, M.R., "Estimation of error rates in discriminant analysis," *Technometrics*, 10, pp.1-11, 1968.

[10] Yanai, H. and Takagi, H., *Handbook of Multivariate analysis,* Gendai-suugakusha, 1986.

[11] Tsukumo, J., "Improved Algorithm for Direction Pattern Matching and Its Application for Hand-printed KANJI Character Classification," *IEICE Technical Report*, PRU90-20, 1990.

[12] Sun, N., Tabara, T., Aso, H. and Kimura M., "Printed Character Recognition Using Directional Element Feature," *IEICE Trans. Syst. & Info.*, vol.J74-D–II, no.3, pp.330-339, Mar., 1991.