8—6

Improvement of Generalization Ability of Kernel-based Fisher Discriminant Analysis for Recognition of Japanese Sign Language Hand Postures, "Yubi-Moji", using K-means method

Katsuya YASUMOTO[*1]
National Institute
of Advanced Industrial Science
and Technology (AIST)

Jun-ya MIZUNO[*2]
Toyohashi University
of Technology

Takio KURITA[*3]
National Institute
of Advanced Industrial Science
and Technology (AIST)

Abstract

Clustering of the training data using K-means method improved the generalization ability of kernel-based Fisher discriminant analysis. In this study, Gaussian type kernel function was used. The proposed method is applied to the vision-based recognition of the 41 Japanese Sign Language (JSL) static hand postures, which express part of the Japanese syllabary. The reflective near-infrared light method was used with few burdens for a hand postures expressioner compared with the conventional methods. The image data obtained by this method can cut a background and include the depth information of the hand as the intensities of image.

1 Introduction

The important role in advanced multi-modal interface is expected to gesture mode. We are paying attention to the sign language as a kind of gesture which has linguistic structure and to recognition of the hand postures of this [1].

The research of recognizing hand postures is roughly divided two ways, a method using a contacted device like a data glove [2], and a method using a noncontact device like a camera [3][4]. As a more natural interface, the latter noncontact vision method is more excellent. For vision-based hand postures recognition, the method of using a silhouette, a color marker, and a color glove to each finger were studied until now. These had the demerit, however, that unnatural simple background were needed, the hand postures in which the outline were similar cannot be discriminated, or a hand had to be colored. In this study, the reflective near-infrared light method without a cumbersome glove-like device is applied to the recognition of the 41 JSL static hand postures. After preprocessing for the raw data, feature vector was extracted by K-means method and Gaussian type kernel function [5]. For classification of 41 postures of 4 subjects, linear discriminant analysis (LDA) was used.

2 Manual Japanese Syllabary, "Yubi-Moji" [6]

"Yubi-Moji" is an element of JSL and expresses the Japanese kana syllabary and number with one hand. **Figure 1** shows the manual Japanese kana syllabary Yubi-Moji. This figure is drawn seen from the listener side. Yubi-Moji of the top row and the leftmost columun in **Fig. 1** are made by reference in manual alphabet of American Sign Language (ASL). While manual alphabet of ASL expresses spelling, these hand postures of JSL is equivalent also to pronunciation (kana is a kind of syllabic). Among Japanese Deaf, Yubi-Moji of the Japanese kana syllabary is mainly used the proper noun with which a sign language word is not decided.

Although there are hand gestures in Yubi-Moji which moves with the form of **Fig. 1** to the right, a top, or body side, that does not treat here. For the more simplification in probrem, in this study, recognition probrem per letter between the 41 letters expressed with a still state in **Fig. 1** are only treated.

[*1]Address: AIST Central 2, 1-1-1, Umezono, Tsukuba, Ibaraki 305-8568, Japan. E-mail: k-yasumoto@aist.go.jp

[*2]Address: Tempaku-cho, Toyohashi, Aichi 441-8500, Japan. E-mail: jjcivic@nrn.tutkie.tut.ac.jp

[*3]E-mail: takio-kurita@aist.go.jp

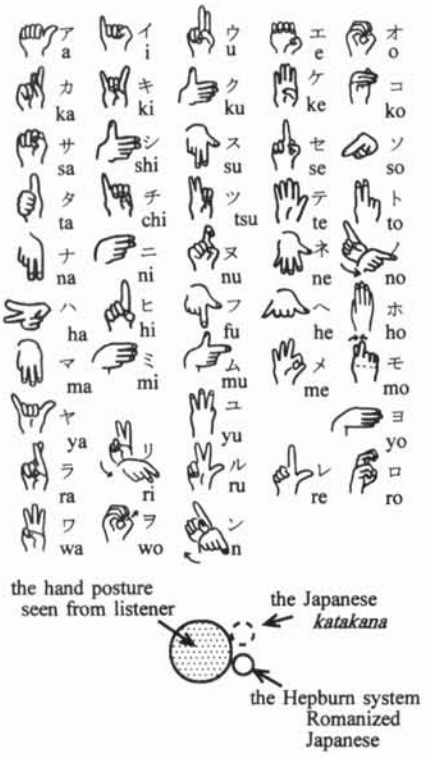

Figure 1. Manual Japanese Syllabary "Yubi-Moji"

3 Data Acquisition [1]

To obtain the raw image data, a device called Motion Processor [7] (photograph in **Figure 2**) was used with a Windows 98 SE personal computer. This device irradiates near-infrared light at object in front of the device, and captures the reflected light by a 64x64 pixel image sensor. Although it depends on near-infrared light reflectance, the reflected light from the background can be cut, because that is weak as compared with that from a near object. As a result, for an object with almost uniform near-infrared light reflectance, the gray scale image which omit background and which include depth information as 256 intensity is obtained. Examples of the image obtained by this device are shown in **Fig. 2, 3**.

Four subjects (males in their 20-40's. hearing persons) expressed repeatedly the Yubi-Moji to the order of the Japanese syllabary in front of the device as shown in **Fig. 2**, and the captured images were recorded as sequential frames on computer. From these data, five frames are extracted as a still image for 41 every letter of every subjects. Among these 5 frames, 3 frames were used as training data and remaining as unknown data. As the whole, we use 492 training data and 328 unknown data frames.

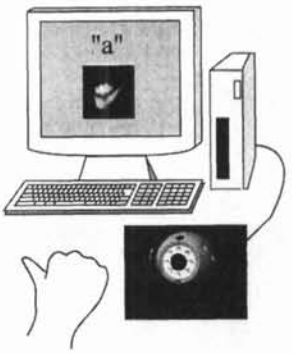

Figure 2. Schematic Arrangement for Data Acquisition

"na" "ra"

Figure 3. Image Example of Hand Postures

The subjects except for a one subject didn't know sign language in advance, so hand postures were taught to them at the time of image acquisition. Since hand posture expression per letter is treated in this study, it is thought that there is no influence skilled in the sign language.

As a preprocessing, the raw data resolution was converted from 64x64 to 32x32 pixels, and the hand image was centered on the frame, and intensity was normalized by linear histogram conversion method. Then, the input vector of 1,024 (=32x32) dimensions which use the intensity of each pixel as an element was obtained by raster scan to every image frame.

4 Feature Extraction by Gaussian kernel and K-means method

The Gaussian type kernel function was used for extracting the feature vector from an input vector.

$$x_i = exp\left(-\frac{d_i^2}{2\sigma^2}\right)$$

is calculated, where d_i denotes distance between input vectors and the i-th training data vector and parameter σ is equivalent to the size of the kernel. This calculation was executed to all training data ($i=1 \sim M$), and the vector $\boldsymbol{x}=(x_1,..,x_M)'$ which makes these x_i an element was obtained as a initial feature vector corresponding to an input vector. If σ is small, the feature vector will consists of local information near input vector. If σ is large, the

feature vector including the information on the large range will be formed.

Although the initial feature vector is obtained, the number of dimensions of the feature vector is the same as the number of traininig data. In this study, since the number of training data is 492, the number of dimensions of the feature vector becomes 492. It is thought possible by reducing this dimensions to raise generalization ability by cutting down the amount of calculation in discriminant analysis and deleting excessive data. Then, before calculating the kernel feature, it tried to compress the number of training data by clustering using the K-mean method. By the K-mean method, the similar data is clustered in the same class, and a representation vector is obtained for every class. These representation vectors are used for the calculation of feature vector instead of a training data vector. Thereby, the dimensions of the feature vector becomes the same as the number of clusters of the K-mean method.

5 Linear Discriminant Analysis [8]

In order to distinguish 41 classes of hand postures, the new feature vector $\boldsymbol{y}=(y_1,..,y_L)^t$ in the discriminant space based on training data is considered. This new feature was obtained by the linear combination of the initial feature vector $\boldsymbol{x}=(x_1,..,x_M)^t$ of the chapter 4 as follows:

$$\boldsymbol{y}=\mathbf{A}^t\boldsymbol{x}$$

where $\mathbf{A}=[a_{ij}]$ is a coefficient matrix and $\mathbf{A}^t$ denotes the transpose of $\mathbf{A}$.

Supposing the set $\{C_k\}_{k=1}^K$ of K classes is given. The within-class and the between-class covariance matrices of the initial features, $\boldsymbol{\Sigma}_W$ and $\boldsymbol{\Sigma}_B$, become

$$\boldsymbol{\Sigma}_W=\sum_{k=1}^{K}\omega_k\boldsymbol{\Sigma}_k,$$

$$\boldsymbol{\Sigma}_B=\sum_{k=1}^{K}\omega_k(\bar{\boldsymbol{x}}_k-\bar{\boldsymbol{x}}_T)(\bar{\boldsymbol{x}}_k-\bar{\boldsymbol{x}}_T)^t$$

where ω_k, $\boldsymbol{\Sigma}_k$, $\bar{\boldsymbol{x}}_k$ are the priori probability, the covariance matrix, the mean vector of Class C_k, respectively, and $\bar{\boldsymbol{x}}_T$ is the total mean vector.

Then, in the discriminant space, the within-class and between-class covariance matrices, $\hat{\boldsymbol{\Sigma}}_W$ and $\hat{\boldsymbol{\Sigma}}_B$, are expressed as

$$\hat{\boldsymbol{\Sigma}}_W=\mathbf{A}^t\boldsymbol{\Sigma}_W\mathbf{A},\qquad \hat{\boldsymbol{\Sigma}}_B=\mathbf{A}^t\boldsymbol{\Sigma}_B\mathbf{A}$$

In order to maximize the discriminating ability of discriminant space, it is required to be the between-class covariance enlarged as much as possible and to be the within-class covariance made as small as possible in discriminant space. Thus, the discriminant criterion

$$J=tr\left(\hat{\boldsymbol{\Sigma}}_W^{-1}\hat{\boldsymbol{\Sigma}}_B\right)$$

is introduced. Obtaining the maximum discliminating ability results in making this J into the maximum. The optimal coefficient matrix which makes a discriminant criterion the maximum is obtained by solving the eigenvalue problem :

$$\boldsymbol{\Sigma}_B\mathbf{A}=\boldsymbol{\Sigma}_W\mathbf{A}\boldsymbol{\Lambda},\quad(\mathbf{A}^t\boldsymbol{\Sigma}_W\mathbf{A}=\mathbf{I})$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix which consists of eigenvalues and $\mathbf{I}$ is a unit matrix. The j-th row of a coefficient matrix $\mathbf{A}$ is the eigenvector corresponding to the j-th size eigenvalue. It is known that in the case of LDA the maximum number of the dimension L of discriminant space is bounded by *min (K-1, M)* .

When unknown image data is given after discriminant space was made from training data as mentioned above, the initial feature vector $\boldsymbol{x}$ is computed by the method of Chapter 4, and new feature $\boldsymbol{y}$ of the unknown data is computed using an above-mentioned coefficient matrix $\mathbf{A}$ obtained from training data.

In this study, to identify the class of the unkown data, we use a simple classifier. The distance between the $\boldsymbol{y}$ of unknown data and the mean $\boldsymbol{y}$ of training data consisted with class C_k was computed. Then it was considered that the unknown data belonged to the class C_k with the nearest distance.

6 Experimental Results

Figure 4 and **Table 1, 2** shows the change of the recognition rate to the training data and that (generalization ability) to the unknown data at the different value of parameter σ and clusters number of the K-means method. It found out that improvement from 97.87 % to 99.09 % in generalization ability was possible by some reducing the number of clusters (492 → 250 ~ 300) , and suitable σ (25). Thus, the method proposed in this study was effective for the improvement of a result.

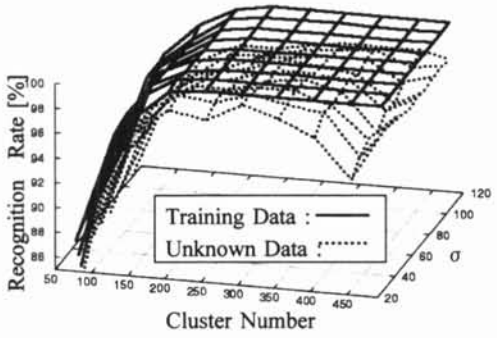

Figure 4 Changes in Recognition Rate vs. kernel size parameter σ and Cluster Number (cluster number 492 = without clustering)

Table 1 Recognition Rate without Clustering

σ	Training Data	Unknown Data
25	100%	97.87%
45	100%	97.56%
65	100%	97.26%
85	100%	97.26%
105	100%	96.95%

Table 2 Recognition Rate using 300 Clustering

σ	Training Data	Unknown Data
25	100%	99.09%
45	100%	98.78%
65	100%	98.48%
85	100%	97.87%
105	100%	97.26%

7 Conclusions

Vision-based recognition of the 41 hand postures of JSL were investigated by using kernel-based Fisher discriminant analysis. The data obtained by reflective near-infrared light method were used. Clustering of the training data using K-means method improved the generalization ability of kernel-based Fisher discriminant analysis.

Although it is still recognition per letter and cannot use for recognition of a word yet, it can use for a command to a robot for example.

Furthermore, application of self-organizing map (SOM) is under examination for recognizing of an anonymous person hand postures, since it is thought that topographic clustering can do SOM [1][9].

A part of this study was supported as Grant-in-Aid for Exploratory Research "14658108" by the Ministry of Education, Culture, Sports, Science and Technology (MEXT).

References

[1] K. Yasumoto, T. Kurita and T. Takahashi : "Vision-based Recognition of the Hand Postures using Self-Organizing Map and Linear Discriminant Analysis", M. Hirose ed., *Human-Computer Interaction INTERACT '01*, pp. 777-778, IOS Press, Ohmsya, 2001.

[2] M. Ishikawa and H. Matsumura : "Recognition of a Hand-Gesture Based on Self-Organization Using a DataGlove", *Proc. of 6th Inter. Conf. on Neural Information Processing*, **II**, pp.739-745, 1999.

[3] V. Pavlovic, R. Sharma and T. S. Huang : " Visual Interpretation of Hand gestures for Human-Computer Interaction: A Review", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, pp.677-695, 1997.

[4] M. V. Lamar, M. S. Bhuiyan and A. Iwata : "Hand Gesture Recognition Using morphological Principal component Analysis and an improved CombNET-II", *Proc. of 1999 IEEE Inter. Conf. on System, Man, and Cybernetics*, Vol. IV, pp.57-62, 1999.

[5] S. Mika, G. Rätsch, J. Weston, B. Schölkopf and K.-R. Müller : "FISHER DISCRIMINANT ANALYSIS WITH KERNELS", in Y. -H. Hu, J. Larsen, E. Wilson and S. Douglas eds., *Neural Networks for Signal Processing* IX, pp. 41-48, IEEE, 1999.

[6] K. Yasumoto :"Sign Language Recognition Research Portal Site (in Japanese)", http://www.neurosci.aist.go.jp/~yasumoto/

[7] TOSHIBA Corp. (1999) : "Development of Motional Interface Device "Motion Processor" ", *TOSHIBA Science and Technology Highlights 1999*, p . 2 (http://www.toshiba.co.jp/tech/review/1999/high99/esub3.htm)

[8] K. Hotta, T. Mishima and T. Kurita : "Scale Invariant Face Detection and Classification Method Using Shift Invariant Features Extracted from Log-Polar Image", *IEICE TRANS. INF. & SYST.*, Vol. E84-D, No. 7, pp.867-878, 2001.

[9] K. Nishida, T. Takahashi and T. Kurita : "A Topographic Kernel-based Regression Method", *Proc. of 6th Joint Conf. on Information Sciences*, pp. 521-524, 2002.