8—19

# Self-organization of viewpoint dependent face representation by the self-supervised learning and viewpoint independent face recognition by the mixture of classifiers

Takio Kurita [1] Hiroyuki Shimai [2] Taketoshi Mishima [2] Takashi Takahashi [3]

[1] Electrotechnical Laboratory

[2] Saitama University

[3] University of Tsukuba (JSPS Research Fellow)

## Abstract

This paper proposes a viewpoint invariant face recognition method in which several viewpoint dependent classifiers are combined by a gating network. The gating network is designed as autoencoder with competitive hidden units. The viewpoint dependent representations of faces can be obtained by this autoencoder from many faces with different views. Multinomial logit model is used for the viewpoint dependent classifiers. By combining the classifiers with the gating network, the network can be self-organized such that one of the classifiers is selected depending on the viewpoint of a given input face image. Experimental results of view invariant face recognition are shown using the face images captured from different viewpoints.

## 1 Introduction

The biological vision system can easily recognize the 3D object from different views. From single-unit recordings in the inferotemporal cortex (IT) after a monkey had learned to recognize a 3D object, it is reported that a small population of neurons indicates remarkable selectivity for individual views of the 3D objects [1]. From the plotting of the responses of such neurons, systematic view-tuning curves for rotations are observed. For some of the tested objects, different neurons are tuned to different views of the same object. Poggio *et al* showed that a simple network, which synthesizes an approximation to a multivariate function representing the object, can achieve viewpoint-invariance by interpolating between a small number of stored views corresponding to an object's training views[2]. A special case of such a network is that of the Radial Basis Functions (RBFs).

For face recognition task, it has also been reported that cells in the primate inferior temporal lobe responds selectively to faces despite substantial changes in viewpoint [3, 4]. Such viewpoint dependent cells have quite broad tuning curves. Perrett *et al*[3] reported broad coding for five principal views of the head: frontal, left profile, right profile, looking up, and looking down. The pose tuning of these cells was on the order of ±40°.

Recently, Bartlett *et al* [5] proposed an attractor network model which can learn viewpoint-invariant face representations from visual experience. Ando *et al* [6] proposed a modular network model which is based on a mixture of non-linear autoencoders. An unsupervised learning algorithm is derived within the framework of the maximum-likelihood estimation. In the learning process, multiple views of 3D objects are randomly presented to the network without providing their object labels. These are suitable for self-organization of viewpoint-invariant representation but not for face classification. For face classification, it is better and straightforward to utilized the information of class labels of the training samples in learning process because they are usually available in training data and these supervised information will improve the recognition performance.

In this paper, we propose a viewpoint invariant face recognition method using mixture of viewpoint dependent classifiers. Several viewpoint dependent classifiers are combined by a gating network. The gating network is designed as autoencoder with competitive hidden units. The viewpoint dependent representations of faces can be obtained by this autoencoder from many faces with different views. Multinomial logit model [7] is used for the viewpoint de-

[1] Address: 1-1-4 Umezono, Tsukuba, Ibaraki 305-8568 Japan. E-mail: kurita@etl.go.jp

[2] Address: 255 Shimo-Okubo, Urawa, Saitama 338-8570 Japan. E-mail: {shimai,mishima}@me.ics.saitama-u.ac.jp

[3] Address: 1-1-1, Tennoudai, Tsukuba, Ibaraki, 305-8573 Japan. E-mail: Takashi@Takahashi.com

pendent classifiers. By combining the classifiers with the gating network, the network can be self-organized such that one of the classifiers is selected depending on the viewpoint of a given input face image. The learning algorithm are presented within a framework of maximum-likelihood estimation. Experimental results of view invariant face recognition are shown using the face images captured from different viewpoints.

## 2 Viewpoint dependent Face Representation

First of all, it is shown that the viewpoint dependent representation of faces can be obtained from many faces with different views by the self-supervised learning. We used a three-layer perceptron (autoencoder) with a soft max function in the hidden layer. The softmax function introduces the competition between the units of the hidden layer.

Let $\boldsymbol{x} = (x_1, \ldots, x_N)^T \in R^N$ denote the input feature vector, $\boldsymbol{g} = (g_1, \ldots, g_H)^T \in R^H$ denote an output vector of the hidden layer and $\boldsymbol{z} = (z_1, \ldots, z_M)^T \in R^{M(=N)}$ denote an output vector of the output layer.

The $h^{th}$ output of the hidden layer is computed as the "softmax" of the weighted sum of the input feature vector $\eta_h = \boldsymbol{v}_h^T \boldsymbol{x}$ as

$$\begin{aligned} g_h &= \frac{\exp(\eta_h)}{1 + \sum_{j=1}^{H-1} \exp(\eta_j)}, \quad h = 1, \ldots, H-1 \\ g_H &= \frac{1}{1 + \sum_{j=1}^{H-1} \exp(\eta_j)}. \end{aligned} \tag{1}$$

The $m^{th}$ output of the output layer is computed by liner function as $z_m = \boldsymbol{w}_m^T \boldsymbol{y}$. The weight vectors $V = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{H-1}\}$ and $W = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{M-1}\}$ can be regarded as the connection weights from the input layer to the hidden layer and from the hidden layer to the output layer in the neural network.

These weights are determined by minimizing the following energy function which is the mean squared error between $\boldsymbol{x}$ and $\boldsymbol{z}$ as

$$E = \frac{1}{2} \sum_{t=1}^{T} ||\boldsymbol{x}_t - \boldsymbol{z}_t||^2. \tag{2}$$

For preliminary experiment, multiple views of face images were randomly presented to this autoencoder. Examples of face images of 10 persons with 25 different views are shown in Figure 1 and 2. The number of training data was 250 (10 classes × 25 directions × 1 image) and the size of feature vector are 2500 (50 × 50). The number of middle layer was set to 3 in this experiment.

Figure 3 shows that the outputs of the hidden layer and the reconstructed faces corresponding to each unit. From these results, we can say that each unit of the hidden layer is properly self-organized and the obtained representation depends on the face views but not on the persons.

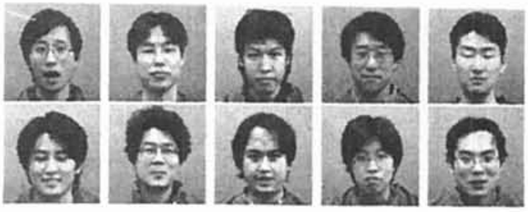

Figure 1: Face images of 10 persons.

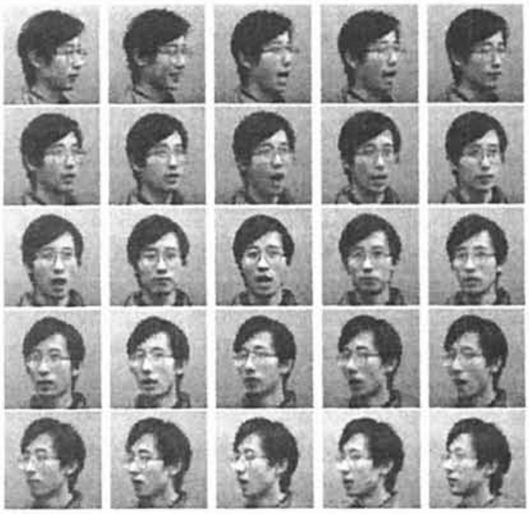

Figure 2: Face images with 25 directions.

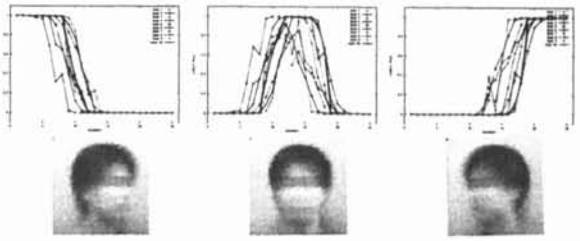

Figure 3: Outputs of the hidden layer and the reconstructed representations.

## 3 Viewpoint independent Face Recognition

The self-organization of view dependent representation suggests the possibility of the view invariant face classification by integrating the viewpoint dependent classifiers with this autoencoder as the gating network. Thus we propose a viewpoint invariant face recognition method using mixture of viewpoint dependent classifiers.

### 3.1 Viewpoint dependent face classifier

The multinomial logit model is used for the viewpoint dependent face classifier. The multinomial logit model is a special case of the generalized linear model [7]. It can be regarded as one of the simplest neural network model for multiway classification problems.

Consider the classification problem with $K$ classes $\{C_1, \ldots, C_K\}$. Let $\boldsymbol{t} = (t_1, \ldots, t_K)^T \in \{0,1\}^K$ denote the binary vector of teacher signal (the desired output of the classifier) with $t_k = 0$ for all $k$ except the correct class $C_j$ with $t_j = 1$. This means that $\sum_{k=1}^K t_k = 1$.

The $k^{th}$ output of the classifier $p_k$ is computed as the "softmax" of the weighted sum of the input feature vector $\eta_k = \boldsymbol{a}_k^T \boldsymbol{x}$. The weight vectors $A = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{K-1}\}$ can be regarded as the connection weights from the input layer to the output layer.

A natural probability model for this classifier is given by

$$P(\boldsymbol{t}|\boldsymbol{x}; A) = \prod_{k=1}^{K} p_k^{t_k}. \tag{3}$$

Then the classifier can be learned by maximizing the log likelihood of (3).

### 3.2 Mixture of Classifiers

For viewpoint invariant face recognition, viewpoint dependent classifiers are combined by using gating network. An example of the architecture is shown in Figure 4. Each classifier receives the same input vector and outputs the classification result as an output vector. The gating network receives the same input as the classifiers and gives the weights of each classifier. The output of the total network is computed as the weighted sum of outputs of the classifiers. This means that the gating network works like a selector of classifiers.

Consider the case there are $H$ viewpoint dependent classifiers. Suppose that the architecture of each classifier is modeled by the multinomial logit model shown in section 3.1. For gating network model, we use the autoencoder shown in section 2. The the weight for the $h^{th}$ classifier $g_h$ is given by the outputs of the hidden layer of the autoencoder shown in (1).

Let the weight vectors $A^{(h)} = \{\boldsymbol{a}_1^{(h)}, \ldots, \boldsymbol{a}_{H-1}^{(m)}\}$ be the connection weights from the input layer to the output layer of the $h^{th}$ classifier and $p_k^{(h)}$ be the $k^{th}$ output of the $h^{th}$ classifier. Then the probability model of the $h^{th}$ classifier is given by $P^{(h)}(\boldsymbol{t}|\boldsymbol{x}; A^{(h)}) = \prod_{k=1}^K p_k^{(h)t_k}$. Thus the probability model for the mixture of $H$ classifiers is given by

$$P(\boldsymbol{t}|\boldsymbol{x}) = \sum_{h=1}^{H} g_h P^{(h)}(\boldsymbol{t}|\boldsymbol{x}; A^{(h)}). \tag{4}$$

By taking the logarithm of both sides, the log likelihood for the mixture of $H$ classifiers is given by

$$l_1 = \log\left[\sum_{h=1}^{H} g_h P^{(h)}(\boldsymbol{t}|\boldsymbol{x}; A^{(h)})\right]. \tag{5}$$

To force the self-organization of viewpoint dependent representations, the additional evaluation function is introduced for the gating network. Note that energy function $E$ can be rewritten into the problem of maximum log likelihood such as

$$l_2 \;=\; -\frac{T}{2\sigma}\sum_{t=1}^{T} ||\boldsymbol{x}_t - \boldsymbol{z}_t||^2, \tag{6}$$

where we suppose that the error $(x_n - z_n)$ is due to the Gaussian distribution $N(0, \sigma^2)$. Then we maximize the following log likelihood

$$L = l_1 + \lambda l_2. \tag{7}$$

Learning algorithm for the mixture of classifiers can be obtained by computing the partial derivatives of this log likelihood. The learning rule for the weight vector of the $h^{th}$ classifier $\boldsymbol{a}_k^{(h)}$ is given by

$$\Delta \boldsymbol{a}_k^{(h)} = s_h(t_k - p_k^{(h)})\boldsymbol{x}. \tag{8}$$

Similarly the learning rule for the weight vector of gating network $\boldsymbol{v}_h$ and $\boldsymbol{w}_m$ are given by

$$\Delta \boldsymbol{v}_h = (s_h - g_h)\boldsymbol{x} + \lambda \sum_{j=1}^{N}(x_j - z_j)(w_{hj} - z_j) y_h \boldsymbol{x} \tag{9}$$

$$\Delta \boldsymbol{w}_m = \lambda(x_m - z_m)\boldsymbol{g} \tag{10}$$

The quantity $s_h$ in these learning rules is defined by

$$s_h = \frac{g_h P_h(\boldsymbol{t}|\boldsymbol{x}; A)}{\sum_{j=1}^{H} g_j P_j(\boldsymbol{t}|\boldsymbol{x}; A)} \tag{11}$$

and this can be considered as the *posterior* probability of the $h^{th}$ gate given input $\boldsymbol{x}$.

## 4 Experiments

To confirm the effectiveness of the proposed face classification method, we have performed experiments using images taken from 25 views. Examples of face

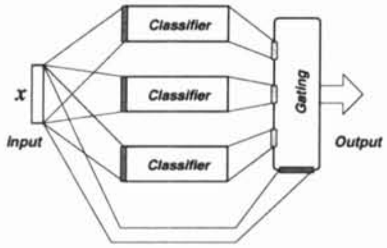

Figure 4: Example of mixture of classifiers

images of 10 persons with 25 different views are shown in Figure 1 and 2. Each image is 50 × 50 pixels with 256 gray levels. The number of training samples is 250 (10 persons × 25 directions × 1 image). Experimental results were evaluated by using 1250 test samples (10 persons × 25 directions × 5 images). From these images, feature vectors are extracted by simply applying Principal Component Analysis (PCA) to the raw image data. In the following experiments, we used the feature vectors whose dimension is 6.

At first, we have performed an experiment using mixture of classifier. The parameter $\lambda$ was set to 0.8. The recognition rate is shown in Table 1 labeled as "Mixture of classifiers ($\lambda = 0.8$)". The outputs of the gating network for each samples are shown in Figure 5.

For comparison, the same learning data is presented to the single classifier (multinomial logit model). The recognition rate is shown in Table 1 labeled as "Single classifier". This result shows that the mixture of classifiers gives better performance than the single classifier.

To investigate the effect of the self-supervised learning, the mixture of classifier with $\lambda = 0$ was trained with the same learning data. The recognition rate is shown in Table 1 labeled as "Mixture of classifiers ($\lambda = 0.0$)". It shows that the performance of the mixture of classifiers with $\lambda = 0$ is little worse than that of the mixture of classifiers with $\lambda = 0.8$. The outputs of the gating network for each samples are shown in Figure 5. These results means that the self-supervised learning is helpful for gating network to learn the viewpoint dependent representations and it causes better recognition results.

Table 1: Recognition rates

| | |
|---|---|
| Mixture of classifiers ($\lambda = 0.8$) | 98.2 |
| Single classifier | 87.2 |
| Mixture of classifiers ($\lambda = 0.0$) | 92.2 |

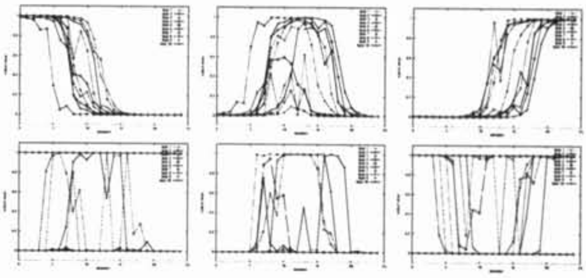

Figure 5: Outputs of gating network: upper row $\lambda = 0.8$, lower row $\lambda = 0.0$

## References

[1] J.Pauls, E.Bricolo, N.Logothetis, "View invariant representations in monkey temporal cortex: position, scale, and rotational invariance," in S.K.Nayer and T.Poggio ed. *Early Visual Learning*, Oxford, 1996.

[2] T.Poggio, S.Edelman, "A network that learns to recognize three-dimensional objects," Nature, Vol.343, pp.263-266, 1990.

[3] D.Perrett, A.Mistlin, A.Chitty, "Visual neurons responsive to faces," Trends Neurosci., Vol.10, pp.358-364, 1989.

[4] M.Hasselmo, E.Rolls, G.Baylis, V.Nalwa, "Object-centered encoding by face-selective neurons in the cortex in the superior temporal sulcus of the monkey," Exp. Brain Res., Vol.75, pp.417-429, 1989.

[5] M.S.Bartlett, T.J.Sejnowski, "Learning viewpoint-invariant face representations from visual experience in an attractor network," Network' Comput. Neural Syst., Vol.9, pp.399-417, 1998.

[6] H.Ando, S.Suzuki, T.Fujita, "Unsupervised visual learning of three-dimensional objects using a modular network architecture," Neural Networks, Vol.12, pp.1037-1051, 1999.

[7] P.McCullaph, J.A.Nelder, "*Generalized Linear Models*," Chapman and Hall, 1983.

[8] J.Kittler, M.Hatef, R.P.W.Duin, J.Matas, "on combining classifiers," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.20, No.3, pp.223-239, 1998.

[9] R.A.Jacobs, M.I.Jordan, S.J.Nowlan, G.E.Hinton, "Adaptive mixtures of local experts,", Neural Computation, Vol.3, pp.79-87, 1991.

[10] M.I.Jordan, R.A.Jacobs, "Hierarchical mixtures of experts and the EM algorithm," Neural Computation, Vol.6, pp.181-214, 1994.