FEATURE EXTRACTION USING SUPERVISED INDEPENDENT COMPONENT ANALYSIS BY MAXIMIZING CLASS DISTANCE

Yoshinori Sakaguchi, Seiichi Ozawa*, and Manabu Kotani***

*Graduate School of Science and Technology, Kobe University, Japan
** Faculty of Engineering, Kobe University, Japan
E-mail: {y-sakag, ozawa}@chevrolet.eedept.kobe-u.ac.jp, kotani@cs.kobe-u.ac.jp

ABSTRACT

Recently, Independent Component Analysis (ICA) has been applied to not only problems of blind signal separation, but also feature extraction of patterns. However, the effectiveness of features extracted by ICA (ICA features) has not been verified yet. As one of the reasons, it is considered that ICA features are obtained by increasing their independence rather than by increasing their class separability. Hence, we can expect that high-performance pattern features are obtained by introducing supervisor into conventional ICA algorithms such that the class separability of features is enhanced. In this work, we propose SICA by maximizing Mahalanobis distance between classes. Moreover, we propose a new distance measure in which each ICA feature is weighted by the power of principal components consisting of the ICA feature. In the recognition experiments, we demonstrate that the better recognition accuracy for two data sets in UCI Machine Learning Repository is attained when using features extracted by the proposed SICA.

1. INTRODUCTION

Recently, Independent Component Analysis (ICA) has been widely known as a decorrelation technique based on high-order moment of input signals [1]. ICA has been so far applied to problems of blind signal separation. On the other hand, feature extraction of images and sounds has been also focused as one of prominent applications of ICA [2,3,4,5].

Bartlett & Sejnowski extracted feature vectors from images of human faces using ICA, and showed that these vectors had greater viewpoint invariance for human faces as compared with Principal Component Analysis (PCA) ones [6]. PCA decorrelates only the second order statistics of input signals, this result indicates that higher-order features are useful for capturing invariant features of face patterns as well as the second-order features. Such invariant characteristics of features extracted by ICA might be attractive for other pattern recognition problems. However, in general, it is not easy to obtain high-performance features using ICA. This might be originated in the fact that class information is not taken into consideration when feature extraction is carried out. Hence, one can expect that good pattern features are obtained by introducing supervisor into conventional ICA algorithms such that the class separability of features is enhanced. Umeyama has applied Supervised ICA by Maximizing Correlation with control signals (here we shall denote as SICA-MC) [7] to images of human faces. In our previous works [8], we have verified the effectiveness of SICA-MC for feature extraction of handwritten characters.

In this paper, we propose a new Supervised ICA by maximizing Mahalanobis Distance between class features (SICA-MD). In Section 2, we describe feature extraction using the conventional ICA algorithm. In Section 3, we propose a Supervised ICA algorithm to extract features such that the class separability of extracted feature is enhanced. In Section 4, we present experimental results for two benchmark data sets in UCI Machine Learning Repository. The last section contains the conclusions of this study.

2. FEATURE EXTRACTION USING ICA

Suppose that we observe a L-dimensional zero mean input pattern at time k, $\boldsymbol{x}(k) = [x_1(k), x_2(k), ..., x_L(k)]^T$, where T means the transposition of matrices and vectors. Assume that this pattern $\boldsymbol{x}(k)$ is linearly composed of M statistically independent components $\boldsymbol{s}(k) = [s_1(k), s_2(k), ..., s_M(k)]^T$ $(M < L)$ as shown bellow:

$$\boldsymbol{x}(k) = \boldsymbol{A}\boldsymbol{s}(k), \tag{1}$$

where $\boldsymbol{A}$ means a $L \times M$ unknown mixing matrix. We estimate $\boldsymbol{A}$ such that each component of $\tilde{\boldsymbol{s}}(k)$ becomes independent each other as much as possible.

In this paper, we shall adopt the bigradient algorithm proposed by Karhunen and Oja as an ICA algorithm [9]. In this algorithm, input patterns $\boldsymbol{x}(k)$ are whitened by PCA at first as shown in the following equation:

$$\boldsymbol{v}(k) = \boldsymbol{D}^{-1/2}\boldsymbol{U}^T\boldsymbol{x}(k), \tag{2}$$

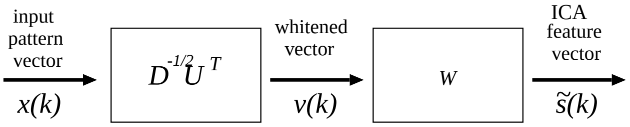

Figure 1: A block diagram of feature extraction of feature vector using ICA.

where $\boldsymbol{D}$ and $\boldsymbol{U}$ are given as follows: $\boldsymbol{D} = \mathrm{diag}[\lambda_1, ..., \lambda_M]$ and $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_M]$. Here, λ_i is the ith largest eigenvalue of the covariance matrix $E\{\boldsymbol{x}(k)\boldsymbol{x}(k)^T\}$ and $\boldsymbol{u}_i$ is the ith eigenvector. Note that these whitening vectors $\boldsymbol{v}(k)$ are mutually uncorrelated and normalized. The uncorrelatedness of features is a necessary condition to be independent. Hence, after the whitening higher order statistical components generally are easily uncorrelated. Using $\boldsymbol{v}(k)$, independent components $\tilde{\boldsymbol{s}}(k)$ can be given by the following equation:

$$\tilde{\boldsymbol{s}}(k) = \boldsymbol{W}\boldsymbol{v}(k), \tag{3}$$

where $\boldsymbol{W}$ is an $M \times M$ separation matrix that is constrained to be an orthonormal matrix. From Eqs. (2)(3), the input-output relation in the ICA process is represented as follows:

$$\tilde{\boldsymbol{s}}(k) = \boldsymbol{W}\boldsymbol{D}^{-1/2}\boldsymbol{U}^T\boldsymbol{x}(k) = \boldsymbol{B}\boldsymbol{x}(k). \tag{4}$$

Fig. 1 shows a block diagram of the information processing in ICA. The ith column vector $\boldsymbol{b}_i$ of $\boldsymbol{B}$ corresponds to a base vector (ICA-base vector) spanning the L-dimensional feature space, and $\tilde{s}_i(k)$ is the projection value for this base vector. Karhunen and Oja has proposed a cost function for ICA that is defined by using the forth-order cumulants of $\tilde{s}_i(k)$, called kurtosis; this cost function is maximized such that $\tilde{s}_i(k)$ is subject to a non-Gaussian distribution. The updating formula of $\boldsymbol{W}$ in the Oja's ICA algorithm is given by the following equation:

$$\begin{aligned}\boldsymbol{W}_{k+1} &= \boldsymbol{W}_k + \mu(\tanh\tilde{\boldsymbol{s}}(k))\boldsymbol{v}(k)^T \\ &\quad +\gamma\boldsymbol{W}_k(\boldsymbol{I}-\boldsymbol{W}_k\boldsymbol{W}_k^T)\boldsymbol{W}_k,\end{aligned} \tag{5}$$

where $\tanh\tilde{\boldsymbol{s}}(k)$ is the vector whose components are calculated by $\tanh\tilde{s}_i(k)$.

In pattern recognition problems, it is more desirable that extracted pattern features belonging to different classes are mutually separated as much as possible in the feature space. ICA algorithm, however, is categorized unsupervised learning; that is, class information is not taken into consideration when feature extraction is carried out. Therefore, high separability of extracted features is not always ensured. To overcome this problem, we shall introduce an additional cost function, which is defined by the Mahalanobis distance between ICA features of two different classes, into the conventional cost function. We call it Supervised ICA by maximizing Mahalanobis Distance between classes (SICA-MD) in the followings.

3. FEATURE EXTRACTION USING SICA-MD

For all sample patterns of class l and m, within-class scatter matrix $\boldsymbol{\Sigma}_{W_{lm}}$ and between-class scatter matrix $\boldsymbol{\Sigma}_{B_{lm}}$ are calculated by the following equations:

$$\boldsymbol{\Sigma}_{W_{lm}} = \sum_{c=l,m} p(c)\frac{1}{n_c}\sum_{v\in\chi_c}(\boldsymbol{v}-\boldsymbol{d}_c)(\boldsymbol{v}-\boldsymbol{d}_c)^T \tag{6}$$

$$\boldsymbol{\Sigma}_{B_{lm}} = \sum_{c=l,m} p(c)(\boldsymbol{d}_c-\boldsymbol{d})(\boldsymbol{d}_c-\boldsymbol{d})^T, \tag{7}$$

$$(l, m = 1, ..., C)$$

where χ_c, n_c, and C mean the pattern set of the class c, the number of samples of class c, and the number of classes, respectively. $\boldsymbol{d}$, $\boldsymbol{d}_c$, and $p(c)$ correspond to the mean vector of all classes, the mean vector of the class c, and the prior probabilities of the class c, respectively.

From all sample patterns of the class l and m, the within-class scatter $\hat{\Sigma}_{W_{lm}}$ and the between-class scatter $\hat{\Sigma}_{B_{lm}}$ of ICA features are given by the following equations:

$$\begin{aligned}\hat{\Sigma}_{W_{lm}} &= \sum_{c=l,m}\Big(p(c)\frac{1}{n_c}\sum_{v\in\chi_c}\boldsymbol{w}_i{}^T(\boldsymbol{v}-\boldsymbol{d}_c)(\boldsymbol{v}-\boldsymbol{d}_c)^T\boldsymbol{w}_i\Big) \\ &= \boldsymbol{w}_i{}^T\Sigma_{W_{lm}}\boldsymbol{w}_i\end{aligned} \tag{8}$$

$$\begin{aligned}\hat{\Sigma}_{B_{lm}} &= \sum_{c=l,m} p(c)\boldsymbol{w}_i{}^T(\boldsymbol{d}_c-\boldsymbol{d})(\boldsymbol{d}_c-\boldsymbol{d})^T\boldsymbol{w}_i \\ &= \boldsymbol{w}_i{}^T\Sigma_{B_{lm}}\boldsymbol{w}_i.\end{aligned} \tag{9}$$

Hence, the Mahalanobis distance $\phi_{lm}(\boldsymbol{w}_i)$ between ICA features of class l and m is given as follows:

$$\phi_{lm}(\boldsymbol{w}_i) = \frac{\hat{\Sigma}_{B_{lm}}}{\hat{\Sigma}_{W_{lm}}} = \frac{\boldsymbol{w}_i{}^T\Sigma_{B_{lm}}\boldsymbol{w}_i}{\boldsymbol{w}_i{}^T\Sigma_{W_{lm}}\boldsymbol{w}_i}. \tag{10}$$

In SICA-MD, the Mahalanobis distance $\phi_{lm}(\boldsymbol{w}_i)$ as well as the independence of features are simultaneously maximized.

As we can see from the example shown in Fig. 2, the separability of ICA features can be enhanced by rotating the obtained ICA base vectors. To do that, we can define an additional cost function such that the weighted sum of Mahalanobis distances between two classes is maximized. Here, we want to separate ICA features of different classes that have similar values. Hence, we shall determine the weights in inverse proportion to the distance between the centers of two classes. More concretely, using the initial vector $\boldsymbol{w}_{i0}$ of $\boldsymbol{w}_i$, the mean features are first calculated from its projection values. Then, their Mahalanobis distances are obtained for

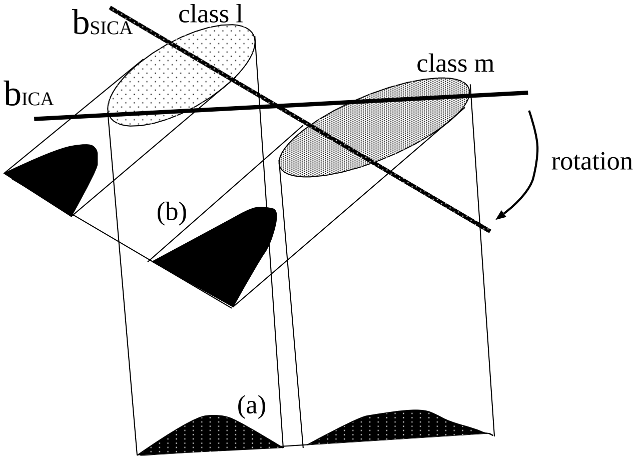

Figure 2: Schematic diagram of the feature transformation by maximizing Mahalanobis distance between ICA features of two different classes. The axis $\boldsymbol{b}_{\mathrm{ICA}}$ is a base vector obtained by unsupervised ICA, and the other axis $\boldsymbol{b}_{\mathrm{SICA}}$ is a base vector obtained by maximizing Mahalanobis distance between ICA features. The two hatched regions schematically mean the distributions of pattern vectors of class l and m. The two black-painted distributions shown in (a) and (b) correspond to the probability density functions of the ICA features projected onto $\boldsymbol{b}_{\mathrm{ICA}}$ and $\boldsymbol{b}_{\mathrm{SICA}}$, respectively.

all combinations of two classes. Here, we denote the distance between mean features of class l and m as $\phi_{lm}(\boldsymbol{w}_{i0})$, and these distances are used for the weights in the cost function. Let us define the weights $1/\phi_{lm}^2(\boldsymbol{w}_{i0})$ for the Mahalanobis distance of $\phi_{lm}(\boldsymbol{w}_i)$ as follows:

$$\Phi(\boldsymbol{w}_i) = \sum_{l=1}^{c-1} \sum_{m=l+1}^{c} \frac{1}{\phi_{lm}^2(\boldsymbol{w}_{i0})} \phi_{lm}(\boldsymbol{w}_i). \quad (11)$$

From this cost function, we can obtain the following derivative $\boldsymbol{\psi}_i$:

$$\boldsymbol{\psi}_i = \frac{\partial \Phi(\boldsymbol{w}_i)}{\partial \boldsymbol{w}_i} = \sum_{l=1}^{c-1} \sum_{m=l+1}^{c} \frac{1}{\phi_{lm}^2(\boldsymbol{w}_{i0})} \frac{\partial(\phi_{lm}(\boldsymbol{w}_i))}{\partial \boldsymbol{w}_i} \quad (12)$$

$$(i = 1, ..., M),$$

where

$$\frac{\partial(\phi_{lm}(\boldsymbol{w}_i))}{\partial \boldsymbol{w}_i} = \frac{2(\{\boldsymbol{w}_i{}^T \Sigma_{W_{lm}} \boldsymbol{w}_i\} \Sigma_{B_{lm}} \boldsymbol{w}_i - \{\boldsymbol{w}_i{}^T \Sigma_{B_{lm}} \boldsymbol{w}_i\} \Sigma_{W_{lm}} \boldsymbol{w}_i)}{\{\boldsymbol{w}_i{}^T \Sigma_{W_{lm}} \boldsymbol{w}_i\}^2}. \quad (13)$$

Equation (13) is added to the Oja's ICA algorithm, and then the update formula of $\boldsymbol{W}$ in the proposed SICA is given as follow:

$$\boldsymbol{W}_{k+1} = \boldsymbol{W}_k + \mu(\tanh \tilde{\boldsymbol{s}}(k)) \boldsymbol{v}(k)^T + \gamma \boldsymbol{W}_k (\boldsymbol{I} - \boldsymbol{W}_k \boldsymbol{W}_k^T) \boldsymbol{W}_k + \alpha \boldsymbol{\Psi}, \quad (14)$$

where

$$\boldsymbol{\Psi} = [\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, ..., \boldsymbol{\psi}_M]^T. \quad (15)$$

Here, α is a positive constant.

4. SIMULATIONS

To investigate the effectiveness of the proposed SICA (SICA-MD), the recognition performance is evaluated for breast-canser-wisconsin data and segmentation data in UCI Machine Learning Repository [10]. The numbers of classes in these problems are two and seven, respectively. The segmentation data are composed of 210 training data and 2100 test data. On the other hand, in the breast-cancer-wisconsin data set, training data and test data are not preliminary designated. Hence, we evaluate the average performance through cross validation. The parameter α in SICA determines the weight of maximizing the separability against the independence. Here, we use some of the training data in order to optimize the parameter α; that is, the recognition performances for various α are evaluated by using these training data, and an optimal value of α is selected. The dimensions of ICA features are set to the same dimensions of PCA features that give the highest recognition performance. From the results of the preliminary experiments, we set the dimensions of breast-cancer-wisconsin data and segmentation data to 8 and 14, respectively.

In order to verify the effectiveness of maximizing the independence and the Mahalanobis distance of features simultaneously, we also evaluate the performance of features extracted by SICA-MD with $\mu = 0$ in Eq. (14), in which only the Mahalanobis distances are maximized. Moreover, as another supervised ICA algorithm, we shall evaluate the performance of features extracted by SICA-MC [7] in which the control signals are set to average patterns for every classes. Since the dimensions of breast-cancer-wisconsin data are 8, a control signal is assigned to four different base vectors to maximize their correlation. On the other hand, since the dimensions of segmentation data are 14, each control signal is assigned to two different base vectors.

In the Oja's ICA algorithm we need to whiten inputs $\boldsymbol{x}(k)$ such that a separation matrix $\boldsymbol{W}$ is constrained to be an orthonormal matrix. Therefore, we do not utilize the information about the eigenvalue λ_i of a PCA base vector $\boldsymbol{u}_i$. This may cause the degradation of the recognition accuracy when ICA features are utilized for recognition. To solve this problem, we shall define the significance of ICA base-vector $\boldsymbol{b}_i$ depending on the eigenvalues λ_j.

From Eq. (4), the relation of the ICA base-vector $\boldsymbol{b}_i$ and the PCA-base vector $\boldsymbol{u}_i$ is given by the following equation:

$$\boldsymbol{b}_i = \frac{w_{i1}}{\sqrt{\lambda_1}}\boldsymbol{u}_1 + ... + \frac{w_{ij}}{\sqrt{\lambda_j}}\boldsymbol{u}_j + ... + \frac{w_{iM}}{\sqrt{\lambda_M}}\boldsymbol{u}_M. \qquad (16)$$

This ICA base vector is normalized as follows:

$$\check{\boldsymbol{b}}_i = (\frac{w_{i1}}{\sqrt{\lambda_1}}\boldsymbol{u}_1 + ... + \frac{w_{ij}}{\sqrt{\lambda_j}}\boldsymbol{u}_j + ... + \frac{w_{iM}}{\sqrt{\lambda_M}}\boldsymbol{u}_M)/C_i, \qquad (17)$$

where

$$C_i = \sqrt{(\frac{w_{i1}}{\sqrt{\lambda_1}})^2 + ... + (\frac{w_{ij}}{\sqrt{\lambda_j}})^2 + ... + (\frac{w_{iM}}{\sqrt{\lambda_M}})^2}. \qquad (18)$$

The similarity of $\check{\boldsymbol{b}}_i$ and $\boldsymbol{u}_j$ is given by $\check{\boldsymbol{b}}_i^T\boldsymbol{u}_j$. Therefore, we can define the following significance ν_i of ICA base-vector $\boldsymbol{b}_i$ based on the similarity to PCA base-vectors $\boldsymbol{u}_j$ that have large $\lambda_j s$:

$$\begin{aligned} \nu_i^2 &= {\lambda_1}^2(\check{\boldsymbol{b}}_1^T\boldsymbol{u}_1)^2 + ... + {\lambda_M}^2(\check{\boldsymbol{b}}_M^T\boldsymbol{u}_M)^2 \\ &= \frac{1}{C_i^2}\{{\lambda_1}^2(\frac{w_{i1}}{\sqrt{\lambda_1}})^2 + ... + {\lambda_M}^2(\frac{w_{iM}}{\sqrt{\lambda_M}})^2\} \\ &= \frac{\lambda_1 w_{i1}^2 + ... + \lambda_M w_{iM}^2}{C_i^2}. \qquad (19) \end{aligned}$$

When the recognition is conducted, ICA feature vectors $\tilde{\boldsymbol{s}}(k)$ are transformed into $\tilde{\boldsymbol{s}}'(k) = [\nu_1 s_1(k), ..., \nu_M s_M(k)]^T$. Then, we calculate the direction cosine between the transformed feature vectors for training data and test data. Let us denote the distances calculated from the ICA feature vectors $\tilde{\boldsymbol{s}}(k)$ and the transformed vectors $\tilde{\boldsymbol{s}}'(k)$ as “distance A”, and “distance B”, respectively. After calculating these distances, the recognition is conducted using a k-nearest neighbor classifier. Here, we set k to 1 and 3.

Table 1 shows the recognition accuracy for segmentation data, and Table 2 shows the recognition accuracy for breast-cancer-wisconsin data. From the results for distance A, we can see that the performance of extracted features except PCA is almost the same. From the results for distance B, we can see that the highest performance features can be extracted by the proposed SICA-MD. Moreover, we can say that the performance of features extracted by SICA-MD is higher than that of features extracted by SICA-MD with $\mu = 0$, in which only class distances are maximized.

In order to examine the independence of features, we evaluate the kurtosis of extracted features, $\tilde{\boldsymbol{s}} = [\tilde{s}_1, ..., \tilde{s}_M]$. The following absolute value of kurtosis is evaluated here:

$$\mathrm{kurt}(\tilde{\boldsymbol{s}}) = \frac{1}{M}\sum_{i=1}^{M}\Big|\frac{E[\tilde{s}_i^4]}{E[\tilde{s}_i^2]} - 3\Big|. \qquad (20)$$

If $\tilde{\boldsymbol{s}}$ has a larger absolute value of kurtosis, one can say that this feature vector is more statistically independent. Table 3

Table 1: Recognition accuracy [%] for segmentation data.

(a)k=1

	distance A	distance B
PCA	79.00	–
ICA	88.05	84.24
SICA-MC	88.00	85.48
SICA-MD	88.05	90.33
SICA-MD ($\mu = 0$)	88.05	87.10

(b)k=3

	distance A	distance B
PCA	73.29	–
ICA	83.19	80.67
SICA-MC	83.19	80.71
SICA-MD	83.19	86.86
SICA-MD ($\mu = 0$)	83.19	83.86

Table 2: Recognition accuracy [%] for breast-canser-wisconsin data.

(a)k=1

	distance A	distance B
PCA	90.00	–
ICA	88.97	89.56
SICA-MC	89.12	89.85
SICA-MD	89.12	90.73
SICA-MD ($\mu = 0$)	89.12	90.44

(b)k=3

	distance A	distance B
PCA	90.44	–
ICA	89.26	90.73
SICA-MC	89.26	90.29
SICA-MD	89.26	90.88
SICA-MD ($\mu = 0$)	89.26	90.73

Table 3: Kurtosis of extracted features for training samples in (a) segmentation data and (b) breast-cancer-wisconsin data.

	(a)	(b)
PCA	20.50	3.045
ICA	26.39	3.396
SICA-MC	19.34	3.056
SICA-MD	15.25	3.071
SICA-MD ($\mu = 0$)	9.52	3.064

Table 4: Class separability of extracted features for training samples in (a) segmentation data and (b) breast-cancer-wisconsin data.

	(a)	(b)
PCA	0.0004	54.77
ICA	0.0033	52.87
SICA-MC	0.031	51.98
SICA-MD	2.278	61.86
SICA-MD ($\mu = 0$)	2.090	61.86

shows the results. We can see that the kurtosis of ICA features is the highest. The reason why the kurtosis of SICA features becomes smaller is that the maximization of class separability as well as the maximization of independence are carried out. Moreover, the kurtosis of features extracted by SICA-MD is larger than SICA-MD with $\mu = 0$. Considering the previous recognition results, we can verify the effectiveness of increasing the independence of features.

We also calculate values of the cost function in Eq.(11) in order to estimate the class separability. Table 4 shows the results of the class separability. We can see that the highest separability is obtained for the ICA features extracted by SICA-MD. This suggests that the term of maximizing separability in Eq.(11) works effectively in the proposed SICA. Moreover, we can see that the class separability of features by extracted SICA-MD is larger than SICA-MD with $\mu = 0$ for segmentation data. From this result, one can say that increasing the independence of features has good influence on the class separability.

5. CONCLUSIONS

The conventional ICA algorithms are categorized into unsupervised learning, and the class information is not taken into consideration when feature extraction is carried out. Therefore, high separability for extracted features is not always ensured. To overcome this problem, we propose a cost function to maximize Mahalanobis distance between class features as well as their independence. Moreover, we propose a new distance measure in which each ICA feature is weighted by the power of principal components consisting of the ICA feature.

From the results of the recognition experiments, it is certified that the features obtained by the proposed supervised ICA is effective as compared with PCA features and the other types of ICA features. From the results on class separability of features, we can see that the maximization of separability works effectively in the proposed SICA. In the future works, we will apply it to other pattern recognition problems, and verify the effectiveness in real world problems.

ACKNOWLEDGEMENT

This research has been supported by the Saneyoshi Scholarship Foundation.

6. REFERENCES

[1] A. Hyvärinen: "Survey on independent component analysis", *Neural Computing Surveys*, **2**, 94-128, 1999.

[2] J. Karhunen, A. Hyvärinen, R. Vigario, J. Hurri, and E. Oja: "Applications of neural blind separation to signal and image processing", *Proc IEEE Int. Conf. Acoust. Speech Signal Process.*, 131-134, 1997.

[3] M. Kotani, Y. Shirata, S. Maekawa, S. Ozawa, and K. Akazawa: "Application of independent component analysis to feature extraction of speech", *Proc of Int. Joint Conf. on Neural Networks* (IJCNN99-Washington DC), CD-ROM #704, 1999.

[4] S. Ozawa, T. Tsujimoto, M. Kotani, and N. Baba: "Application of independent component analysis to hand-written Japanese character recognition", *Proc of International Joint Conf. on Neural Networks* (IJCNN99-Washington DC), CD-ROM #462, 1999.

[5] Y. Watanabe, M. Hiratani, and T. Nagano: "Feature extraction of palm prints using supervised independent component analysis", *CD-ROM Proc. of 7th Int. Conf. on Neural Info. Processing*, 2000.

[6] M. S. Bartlett, H. M. Lades and T, and J. Sejnowski: "Independent component representations for face recognition", *Proc. of the SPIE*, **3299**, 528-539, 1997.

[7] S. Umeyama, S. Akaho, and Y. Sugase: "Supervised independent component analysis and its applications to face image analysis" (in Japanese), *Tech. Report of IEICE*, **NC 99-2**, 9-16, 1999.

[8] S. Ozawa, Y. Sakaguchi, and M. Kotani: "A study of feature extraction using supervised independent component analysis", *Proc. of Int. Conf. on Neural Networks* (IJCNN2001-Washington DC), 2958-2963, 2001.

[9] J. Karhunen and E. Oja; "A class of networks for independent component analysis", *IEEE Trans.on Neural Networks*, **8**, 3, 486-503, 1997.

[10] http://www.ics.uci.edu/~mlearn/MLRepository.html