Discriminant Independent Component Analysis

Chandra Shekhar Dhir and Soo-Young Lee, *Member, IEEE*

Abstract—**A conventional linear model based on Negentropy maximization extracts statistically independent latent variables which may not be optimal to give a discriminant model with good classification performance. In this paper, a single-stage linear semisupervised extraction of discriminative independent features is proposed. Discriminant independent component analysis (*d*ICA) presents a framework of linearly projecting multivariate data to a lower dimension where the features are maximally discriminant with minimal redundancy. The optimization problem is formulated as the maximization of linear summation of Negentropy and weighted functional measure of classification. Motivated by independence among extracted features, Fisher linear discriminant is used as the functional measure of classification. Experimental results show improved classification performance when *d*ICA features are used for recognition tasks in comparison to unsupervised (principal component analysis and ICA) and supervised feature extraction techniques like linear discriminant analysis (LDA), conditional ICA, and those based on information theoretic learning approaches. *d*ICA features also give reduced data reconstruction error in comparison to LDA and ICA method based on Negentropy maximization.**

Index Terms—**Discriminant independent component analysis, feature extraction, Fisher's linear discriminant, negentropy.**

I. Introduction

FAST computation involved in a real-time system is severely challenged by the dimensionality of raw data. High-dimensional multivariate data not only imposes computational constraints but also makes the problem ill posed due to high redundancy [1]–[4]. Unsupervised feature extraction method decomposes data into its basis and features (latent variables) with reduced redundancy. Among several unsupervised frameworks, we concentrate on principal component analysis (PCA) and independent component analysis (ICA) in this paper.

PCA tries to find orthonormal representation of data which minimizes the squared error between the original data and its reconstruction [5], [6]. ICA is a multivariate approach of data representation which extracts statistically independent features [7]–[11]. The independence among the low-dimensional representations of data has made ICA an attractive algorithm for application in the field of blind source separation and feature extraction. Several approaches have been used to extract ICA features among which Infomax [8] and Negentropy maximization [7] are the most popular. While, the Infomax algorithm maximizes the entropy of the nonlinearly transformed outputs of the ICA network, Negentropy-based ICA derives its learning rule by maximizing the sum of marginal Negentropy of the outputs of ICA network. We used gradient-based Negentropy maximization algorithm to extract features with minimum statistical dependence which are constrained to have unit variance [7]. Another advantage of ICA based on the Negentropy approach is that it can extract independent feature subspace with cardinality less than the input data dimension.

However, a feature that is good for data representation as in the case of PCA and ICA may not necessarily be good for classification performance, as it does not utilize the association between the observations and its class. Pattern classification which utilizes class information can be broadly categorized as the wrapper method [12], filtering method, and supervised feature extraction method.

Feature selection using the wrapper method is known to suffer from exhaustive combinatorial search problem. Although the wrapper method gives good recognition performance for the given classifier, its lack of generality to an unknown classifier makes it a less favorable choice. On the other hand, filtering method is a suboptimal greedy approach which gives each feature a score based on certain discriminant criteria (Fisher criteria, mutual information (MI), etc.) [13]–[19]. Meaningful features can be selected depending on the relevance of the score to classification accuracy.

Fisher criterion is a suboptimal feature selection method which assumes that the conditional distribution of the feature given class is Gaussian. This $apriori$ assumption may not result in an optimal selection of a feature having non-Gaussian conditional distribution. On the other hand, MI between features and a given feature class has been widely used as a feature selection criterion. It attempts to maximize the relevance (importance to classification task) and minimize the redundancy among the features (mutually independent features) [14]–[17]. However, estimation of MI suffers from the curse of dimensionality. Also, feature selection using MI criteria becomes mathematically intractable when dependency of the feature to be selected is considered with a large set of other features. Filtering approach ensures that the subset of selected features maintains the statistical properties of unsupervised features. However, it would still be nice if we had only one stage where extraction and selection can take place simultaneously. In another method, Huang *et al.* extracted

Manuscript received February 22, 2010; revised November 18, 2010; accepted February 13, 2011. Date of publication April 25, 2011; date of current version June 2, 2011. This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea and the Ministry of Education, Science and Technology, under Program 2009-0092812 and Program 2010-0028722.

C. S. Dhir was with the Department of Bio and Brain Engineering, Brain Science Research Center, Korea Advanced Institute of Science and Technology, Daejeon 305-701, Korea. He is now with LG Electronics, Seoul 150-721, Korea (e-mail: shekhardhir@gmail.com).

S.-Y. Lee is with the Department of Electrical Engineering, Brain Science Research Center, Korea Advanced Institute of Science and Technology, Daejeon 305-701, Korea (e-mail: sylee@kaist.ac.kr).

Digital Object Identifier 10.1109/TNN.2011.2122266

1045–9227/$26.00 © 2011 IEEE

independent features and later performed classification through an optimal scoring algorithm [18].

Linear discriminant analysis (LDA) is a single-stage supervised feature extraction technique which finds linear projections of data so that the classes are maximally separable [3], [20]–[22]. Other classical methods such as sliced inverse regression [23] and canonical correlation analysis (CCA) [24] also attempt to maximize the discriminative ability of extracted features. SVM-2K proposed by Farquhar *et al.* presents an alternative single stage learning approach which combines the two-stage kernel CCA with support vector machines (SVMs) [25]. In this paper, we have concentrated on classification experiments using LDA features as a baseline.

For high-dimensional multivariate data with small samples, LDA is prone to singularity problems. In an attempt to overcome the small sample size problem, PCA is used to reduce the dimension of original data and then LDA is performed on these extracted orthonormal features. As an extension to this idea of Fisherfaces [26], Fei *et al.* proposed a two-stage approach where data is initially projected to a low-dimensional space of independent features and later LDA is performed [27]. Projection of independent features to a discriminant feature space using LDA does not ensure that the features are still independent.

In an attempt to overcome the parametric assumption of LDA, information theoretic learning (ITL) has been proposed which gives a generative discriminant model. ITL approaches to linear supervised feature extraction uses the concept of Reyni's entropy [28] to maximize the quadratic mutual information (QMI) of transformed data and classes. It uses nonparametric estimates of a feature and its conditional distribution [29]–[31]. The density functions are estimated using Parzen window method with Gaussian kernel. Although the use of the Gaussian kernel simplifies the update rules making the calculation of QMI computationally scalable, it suffers from the curse of dimensionality. There is also no theoretical explanation which claims that the maximization of QMI is equivalent to the maximization of mutual information (MMI) using Shannon's entropy [31].

As an alternative approach, the objective of MMI between multivariate features and class variables can be approximated by maximization of the marginal sum of MI between a feature and class ignoring MI among the features [32]. In an elegant manner, Murillo *et al.* showed that the MI between a feature and class can be defined using a Negentropy-based contrast function. It also showed improved recognition performance over feature extraction method, which maximizes the QMI. However, maximization of marginal sum of MI between a feature and class inherently increases the redundancy among the features and minimizes the redundancy between features given class. An appropriate linear transform should maintain statistical independence among the features and at the same time enhance the discriminative property of the features.

Conditional independent component analysis (CICA) is another approach that extends the ICA framework to supervised learning [33]. It minimizes the Kullback–Liebler (KL) divergence between the joint and the product of marginal conditional distribution of the outputs. The dual objective function of CICA shows similarity with the information bottleneck method [34] which extracts minimally redundant features that maximally preserve the information about auxiliary data (i.e., class information). Independently, Amato *et al.* had proposed independent discriminant component analysis which is similar to CICA [35]. In another approach, bilinear discriminant component analysis finds projections with optimal class discrimination such that projected features make independent contribution to the discrimination [36].

In this paper, a single-stage semisupervised approach for extraction of discriminative independent features is proposed. This *d*ICA method maximizes the FLD along with the sum of marginal Negentropy of the separated independent features. Thus, *d*ICA is an attempt to combine representative and discriminant model for good classification. It also finds representative latent variables (features) with minimal statistical dependencies among them. Recognition results show improved classification performance when the proposed *d*ICA features are used over PCA, ICA, LDA, CICA, and ITL based features. It also shows better data representation in comparison to ICA and LDA, giving lesser data reconstruction error. In comparison to the two-stage approach proposed by Fei *et al.* [27], *d*ICA is advantageous as it simultaneously extracts discriminant and independent features with reduced computational complexity.

The remainder of this paper is organized as follows, Section II reviews the gradient-based learning rule for maximization of Negentropy. In Section III, *d*ICA is presented. Experimental setup and results are presented in Section IV. In Section V, advantages of *d*ICA features for classification and data reconstruction are discussed. Nonsingular *d*ICA is introduced in Section VI, which avoids singularity problems and finds an optimal weight for the discriminant function. This is followed by conclusions in the last section.

II. Criterion for Independent Features: Negentropy Maximization

A low-dimensional representation of multivariate data using independent features can be achieved by maximizing Negentropy of the extracted features [7]. Given N samples of a P-dimensional multivariate observation, $\mathbf{X}_{P\times N} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$, ICA based on Negentropy maximization finds mutually independent low-dimensional features.

The task is to linearly decompose the whitened observation $\mathbf{Z}_{L\times N} = \hat{\mathbf{W}}\mathbf{X}$: $L \leq P$, into a basis matrix $\mathbf{A}$ and a feature matrix $\mathbf{Y}_{R\times N} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_R]^T$: $R \leq L$ such that the features $\mathbf{y}_i$ are statistically independent. Here, L is the number of principal axes selected from data $\mathbf{X}$ and R is the number of independent sources extracted from $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N]$. $\hat{\mathbf{W}}$ is the whitening matrix which is given as

$$\hat{\mathbf{W}} = \mathbf{D}^{-0.5}\mathbf{U}^T \tag{1}$$

where $\mathbf{D}$ is a $L \times L$ diagonal matrix whose ith element is the ith largest eigenvalues of the covariance matrix of $\mathbf{X}$, i.e., $d_{ii} = \lambda_i$. $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_L]$, where $\mathbf{u}_i$ is the eigenvector corresponding to the eigenvalue λ_i.

The extracted independent features $\mathbf{Y}$ are given as

$$\mathbf{Y} = \mathbf{W}^T\mathbf{Z} = \mathbf{W}^T\hat{\mathbf{W}}\mathbf{X} \tag{2}$$

where $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_R]$ and $\mathbf{w}_i$ is an L-dimensional column vector.

A. Approximation of Negentropy of a Random Vector

Negentropy is a good statistical measure of non-Gaussianity of a random variable. An approximation of the marginal Negentropy of $\mathbf{y}_i$ can be given as [7]

$$J(\mathbf{y}_i) \approx \kappa_1 \left(E\left(G^1(\mathbf{y}_i)\right)\right)^2 + \kappa_2 \left(E\left(G^2(\mathbf{y}_i)\right) - E\left(G^2(\boldsymbol{\nu})\right)\right)^2 \tag{3}$$

where $G^1(\cdot)$ and $G^2(\cdot)$ are non-quadratic odd and even functions, respectively. Negentropy approximation in (3) is often more accurate and robust than the cummulant-based approximations. For robust estimation, these functions should not grow faster than $\mathbf{y}_i^2$ as $|\mathbf{y}_i|$ varies. Some of the popular choices of G^1 and G^2 for random vector $\mathbf{y}_i$ with symmetric distribution are as follows:

$$\begin{aligned} G^1(\mathbf{y}_i) &= 0 & (4)\\ G^2(\mathbf{y}_i) &= \frac{1}{a_1}\log\cosh a_1\mathbf{y}_i, \quad 0 < a_1 \leq 1 & (5)\\ G^2(\mathbf{y}_i) &= -\exp\left(-\frac{\mathbf{y}_i^2}{2}\right). & (6) \end{aligned}$$

Assuming that the distribution of $\mathbf{y}_i$ is symmetric $E(\mathbf{y}_i^3) = 0$, one may further simplify the Negentropy approximation in (3) to

$$J(\mathbf{y}_i) = \kappa_2 \left(E\left(G(\mathbf{y}_i)\right) - E\left(G(\boldsymbol{\nu})\right)\right)^2 \geq 0 \tag{7}$$

where G^2 is replaced by G for simplification and the value of $\kappa_2 = 24/16\sqrt{3} - 27$. ν represents a univariate Gaussian distribution with the same mean and standard deviation as of $\mathbf{y}_i$.

B. Extraction of Independent Features by Negentropy Maximization

Considering symmetric distribution of extracted features, an attempt is made to maximize the sum of marginal Negentropy of the outputs $\mathbf{y}_i$. In addition, the covariance of zero mean $\mathbf{Y}$ is constrained to identity matrix, $\mathbf{I}_{R\times R}$, i.e., $E\left(\mathbf{Y}\mathbf{Y}^T\right) = \mathbf{I}$. Constraining the covariance of $\mathbf{Y}$ to $\mathbf{I}$ results in orthonormal $\mathbf{W}$.

Maximization of sum of marginal Negentropy of $\mathbf{y}_i$ along with the unit covariance constraint of $\mathbf{Y}$ can be modeled as a constrained Lagrangian equation, which is given as [37]

$$\begin{aligned} \tilde{L}(\mathbf{W}) = &\sum_{i=1}^{R}\left[E\left(G\left(\mathbf{w}_i^T\mathbf{Z}\right)\right) - E\left(G\left(\nu\right)\right)\right]^2 \\ &+ \sum_{i=1}^{R}\beta_i\left(\mathbf{w}_i^T\mathbf{w}_i - 1\right). \end{aligned} \tag{8}$$

Gradient-based adaptation of $\mathbf{w}_i$ along with symmetric orthogonalization is given by the following learning rule:

$$\begin{aligned} \Delta\mathbf{w}_i &= \eta\left(\gamma_i E\left(\mathbf{Z}g\left(\mathbf{w}_i^T\mathbf{Z}\right)\right) + 2\beta_i\mathbf{w}_i\right) & (9)\\ \mathbf{W} &\leftarrow \left(\mathbf{W}\mathbf{W}^T\right)^{-\frac{1}{2}}\mathbf{W} & (10) \end{aligned}$$

where $\gamma_i = 2\left[E\left(G\left(\mathbf{w}_i^T\mathbf{Z}\right)\right) - E\left(G\left(\nu\right)\right)\right]$ and η is the learning rate. $g(\cdot)$ is the first-order differential of $G(\cdot)$.

Choice of the even function G as in (6) can be used to analytically estimate the value of γ_i as

$$\gamma_i = 2\left(-\frac{\sum_{n=1}^{N}\exp\left(-\frac{y_{in}^2}{2}\right)}{N} + \frac{1}{\sqrt{2}}\right) \tag{11}$$

where y_{in} is the nth sample of the ith feature. At convergence, the value of constrained Lagrangian given by (8) is $\sum_{i=1}^{R}\gamma_i^2$.

The value of β_i can be found by setting the first-order differential of (8) with respect to $\mathbf{w}_i$ equal to 0 which is given as

$$\beta_i = -\frac{1}{2}\gamma_i E\left(\mathbf{y}_i g\left(\mathbf{y}_i\right)\right). \tag{12}$$

Features extracted by maximization of the objective function given in (8) results in maximally non-Gaussian features. Since some features may be common to all classes, the feature set may not ensure good classification performance. In the next section, we present dICA approach which exploits class label to extract discriminant features with reduced higher order statistical dependence.

III. DICA

In this section, a single-stage semisupervised approach to extract discriminant independent features is presented. Classification performance can be improved by exploiting *apriori* information of the class labels. In addition to the maximization of sum of Negentropy of $\mathbf{y}_i$ as discussed in Section II-B, we also maximize the functional measure of classification performance for the given features.

The new optimization is a dual maximization problem which jointly maximizes the sum of marginal Negentropy and the functional measure of classification of the features, $\mathbf{Y}$, under the unit covariance constraint of $\mathbf{Y}$.

The constrained dual maximization problem can be defined by the Lagrangian, which is given as

$$\hat{L}\left(\mathbf{W}\right) = \tilde{L}\left(\mathbf{W}\right) + k\phi(\mathbf{W}, \mathbf{Z}, C) \tag{13}$$

where k is a constant multiplier and $\phi(\mathbf{W}, \mathbf{Z}, C)$ is the functional measure of classification performance of features $\mathbf{Y}$ given C.

Adaption of $\mathbf{w}_i$ can be achieved by using gradient ascent along the function given in (13). The learning rule is given as

$$\begin{aligned} \Delta\mathbf{w}_i &= \eta\left(\gamma_i(E\left(\mathbf{Z}g\left(\mathbf{w}_i^T\mathbf{Z}\right)\right) + k\tfrac{\partial\phi(\mathbf{W},\mathbf{Z},C)}{\partial\mathbf{w}_i} + 2\beta_i\mathbf{w}_i\right) & (14)\\ \mathbf{W} &\leftarrow \left(\mathbf{W}\mathbf{W}^T\right)^{-\frac{1}{2}}\mathbf{W}. & (15) \end{aligned}$$

The value of β_i can be obtained from the stationary point solution of (13) as

$$\beta_i = -\frac{1}{2}\left(\gamma_i E\left(\mathbf{y}_i g\left(\mathbf{y}_i\right)\right) + k\mathbf{w}_i^T \frac{\partial \phi(\mathbf{W}, \mathbf{Z}, C)}{\partial \mathbf{w}_i}\right). \qquad (16)$$

A. Functional Measure of Classification Performance: Fisher's Linear Discriminant

1) LDA: LDA linearly maps high-dimensional data to a lower dimension, $f : \mathbb{R}^{L\times N} \rightarrow \mathbb{R}^{R\times N}, R < L$, such that the features are discriminant. An optimal linear transformation $\mathbf{Y} = \tilde{W}^T\mathbf{Z}$ can be obtained by maximizing the FLD, which is given as [3], [22]

$$\tilde{W} = \arg\max_{\mathbf{W}} F(\mathbf{W}) = \arg\max_{\mathbf{W}} \frac{\left|\mathbf{W}^T\mathbf{S}_B\mathbf{W}\right|}{\left|\mathbf{W}^T\mathbf{S}_W\mathbf{W}\right|} \qquad (17)$$

where $|\cdot|$ gives the determinant of the input matrix.

The between-class scatter matrix $\mathbf{S}_B$ and within-class scatter matrix $\mathbf{S}_W$ are given as

$$\mathbf{S}_B = \frac{1}{N}\sum_{c=1}^{C} N_c\left(\mathbf{m}_c - \mathbf{m}\right)\left(\mathbf{m}_c - \mathbf{m}\right)^T \qquad (18)$$

$$\mathbf{S}_W = \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{z}'_n - \mathbf{m}_{c(n)}\right)\left(\mathbf{z}'_n - \mathbf{m}_{c(n)}\right)^T \qquad (19)$$

respectively. Here N_c is the number of samples in class c, and $c(n)$ is the class index of sample n. $\mathbf{m}_c = [E(\mathbf{z}_1|c), \ldots, E(\mathbf{z}_L|c)]^T$ is a column vector of conditional mean of $\mathbf{Z}$ given c, and $\mathbf{m} = [E(\mathbf{z}_1), \ldots, E(\mathbf{z}_L)]^T$ is a column vector defined by the marginal mean of $\mathbf{Z}$. $\mathbf{z}'_n$ is the nth column vector of $\mathbf{Z}$. Since the rank of $\mathbf{S}_B$ is $(C-1)$, it is theoretically possible to find upto $R = (C-1)$ features. It can be easily shown that the projections of $\mathbf{Z}$ using $\tilde{W}$ are the eigenvectors of $(\mathbf{S}_W)^{-1}\mathbf{S}_B$.

The solution of the maximization problem given in (17) for LDA can obtain only $(C-1)$ optimal projections of multivariate data. Several researchers have proposed extensions of LDA which can obtain more than $(C-1)$ features. In this paper, LDA features are obtained using the subspace search method proposed by Murillo *et al.* [32]. It searches for a determined number of projections in the complementary space of already obtained features using Gram–Schmidt orthonormalization.

2) Discriminant Analysis with Independent Features in $\mathbb{R}^{R\times N}$: The dICA feature extraction method attempts to maximize the discriminant function and at the same time makes the output features statistically independent. Using the additive property of statistically independent features, the discriminant function can be defined as the sum of the logarithm of $F(\mathbf{w}_i)$ [22]

$$\phi(\mathbf{W}, \mathbf{Z}, C) = \log F(\mathbf{W}) = \sum_{i=1}^{R}\log F(\mathbf{w}_i). \qquad (20)$$

In the feature space $\mathbf{Y}$, the discriminant function is given as

$$\begin{aligned}\phi(\mathbf{W}, \mathbf{Z}, C) &= \sum_{i=1}^{R}\log\frac{\mathbf{w}_i^T\mathbf{S}_B\mathbf{w}_i}{\mathbf{w}_i^T\mathbf{S}_W\mathbf{w}_i} \\ &= \sum_{i=1}^{R}\log\frac{\sum_{c=1}^{C}N_c\left(\mu_{ic}-\mu_i\right)^2}{\sum_{c=1}^{C}N_c\sigma_{ic}^2} \qquad (21)\end{aligned}$$

$$\mu_{ic} = \frac{1}{N_c}\sum_{n\in Class\ c} y_{in}, \quad \sigma_{ic} = \frac{1}{N_c}\sum_{n\in Class\ c}\left(y_{in} - \mu_{ic}\right)^2 \qquad (22)$$

where C is the total number of classes, $\mu_{ic} = E(\mathbf{z}_i|c)$ is the mean of ith feature given class c, and N_c is the number of samples corresponding to class c. $\mu_i = E(\mathbf{z}_i)$ is the mean of ith feature.

$F(\mathbf{w}_i)$ is the measure of a given feature $\mathbf{y}_i$ to correctly associate data with its classes. To maximize (21), gradient-based update rule of $\mathbf{W}$ requires calculation of partial derivative of $\phi(\mathbf{W}, \mathbf{Z}, C)$ with respect to w_{li} (refer to the Appendix for detailed derivation)

$$\frac{\partial\phi(\mathbf{W}, \mathbf{Z}, C)}{\partial w_{li}} = 2\sum_{c=1}^{C}\sum_{n\in Class\ c}\left(A_{ic} - B_{ic}\right)z_{ln} \qquad (23)$$

$$A_{ic} = \frac{\left(\mu_{ic} - \mu_i\right)}{\sum_{c'=1}^{C}N_{c'}\left(\mu_{ic'} - \mu_i\right)^2}, \quad B_{ic} = \frac{\left(y_{in} - \mu_{ic}\right)}{\sum_{c'=1}^{C}N_{c'}\sigma_{ic'}^2} \qquad (24)$$

where z_{ln} is the nth sample of lth whitened observation. In the absence of additional control parameters, the logarithm in (21) is beneficial with simple derivatives. The logarithm function also discourages overlap among the classes more seriously, i.e., if the interclass mean is smaller than the within class variance, the logarithm of $F(\mathbf{w}_i)$ is negative.

IV. Experimental Results

To evaluate the performance of dICA presented in Section III, recognition experiments were performed on several publicly available datasets. Most of the datasets were obtained from University of California at Irvine (UCI) repository [38]. The recognition performance of dICA was compared with PCA, LDA, and ICA using Negentropy maximization. We also compared the recognition performance of dICA with MMI [32], maximization of quadratic mutual information (MQMI) [31], and CICA [33], which are discussed below.

1) MMI [32]: Murillo *et al.* defined the objective function as the sum of marginal MI between $\mathbf{y}_i$ and C, which is given as

$$J^{MMI}(\mathbf{Y}) = \sum_{i=1}^{R} I(\mathbf{y}_i; C). \qquad (25)$$

The task is to find the optimal projection matrix $\mathbf{W}$ that maximizes the objective function in (25) under the unit covariance constraint of $\mathbf{Y}$

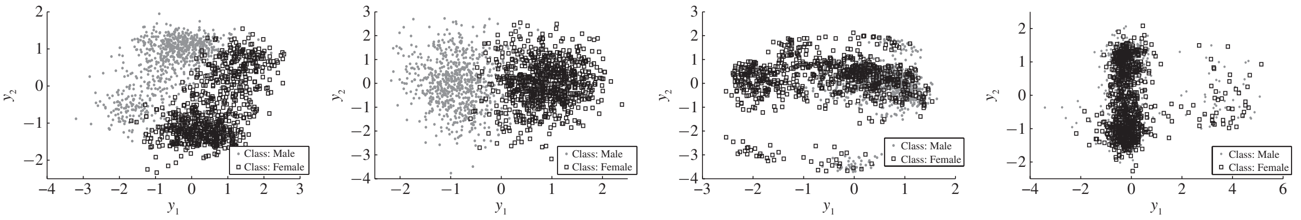
$$\mathbf{W} = \arg\max_{\mathbf{W}} J^{MMI}(\mathbf{W}^T\mathbf{Z}). \qquad (26)$$

2) MQMI [31]: ITL-based linear feature extraction method finds projection of data that maximizes the QMI between the features and class variables. In [31], the transformation matrix $\mathbf{W}$ is constrained to be a rotation matrix which reduces the number of parameters to be estimated to $R \times (L - R)$ instead of $R \times L$ as in case of [29].

3) **CICA** [33]: CICA minimizes the KL divergence between the joint and the product of marginal conditional distribution of the outputs. The task is to find an optimal

TABLE I
DATASET USED FOR EXPERIMENTAL STUDY

Database	Number of classes	Dimen- -sion	Training samples	Test/ Validation samples	Number of Cross validation
UCI Handwritten digits [40]	10	64	3823	1797	
German emotion speech (AIBO Features) [41], [42]	7	200	424	111	5
Postech faces gender recognition [43]	2	22 500	1512	216	4 random test experiments
Letter recognition dataset [44]	26	16	16 000	4000	
Multiple feature dataset [45]	10	76	1500	500	4
Isolated spoken letter recognition	26	616	4677	1559	4



(a) (b) (c) (d)

Fig. 1. 2-D scatter plot of features obtained from the training samples of gender recognition using the Postech face database. (a) dICA($k = 1$). (b) LDA. (c) PCA. (d)ICA.

W that minimizes the objective function [33]

$$J^{CICA}(\mathbf{W}) = D_{KL}\left(p(\mathbf{Y}, C)\,||p(C)\prod_{i=1} p(\mathbf{y}_i, C)\right). \tag{27}$$

It can be shown that, for an invertible transformation, the above objective function can be rewritten as maximization of

$$J^{CICA}(\mathbf{W}) = \sum_{i=1}^{L} J(\mathbf{y}_i) + \sum_{i=1}^{L} I(\mathbf{y}_i; C). \tag{28}$$

CICA is similar to dICA approach when $\phi(\mathbf{W}, \mathbf{Z}, C) = J^{MMI}(\mathbf{Y})$ for $k = 1$. In comparison with MMI approach, CICA also attempts to minimize the redundancy among the linear projection of data. To extract $R \leq L$ features from the whitened data, we maximize (25) under the unit covariance constraint of $\mathbf{Y}$.

The datasets mentioned in Table I is used in the experimental study. They are centered and projected to an L-dimensional orthonormal space using whitening matrix $\hat{\mathbf{W}}$. The value of L for the different datasets is given in Table II. In case of gender recognition problem using Postech faces, we extracted atmost 20 features from the 50 principal components.

A. Classifier

In this paper, SVM with Gaussian kernel is used as the classifier [47]. Publicly available implementation of SVM, i.e., SVM^{Light}, is used in all experiments [48]. SVM classifier is theoretically well established with global solutions and is less sensitive to the curse of dimensionality. The two main parameters, i.e., soft margin χ and the standard deviation of Gaussian kernel σ, are varied and the best classification performance in the $\chi - \sigma$ parameter space is chosen. χ is a regularization parameter that controls the tradeoff between the margin and the training error. In total, 36 SVMs were trained with $\chi = \{0.5, 1, 3, 5, 10, 20\}$ and $\sigma = \{2, 5, 10, 20, 25, 30\}$. The best recognition performance for a given number of extracted features is the maximum recognition performance among all the SVMs.

In addition, nearest neighbor (NN) classifier with cosine similarity metric between training and test features is also considered. It has been shown that NN classifier shows good classification performance provided an appropriate linear transformation is performed on the data [49]. In this paper, we use 1-NN classifier.

B. Experimental Results

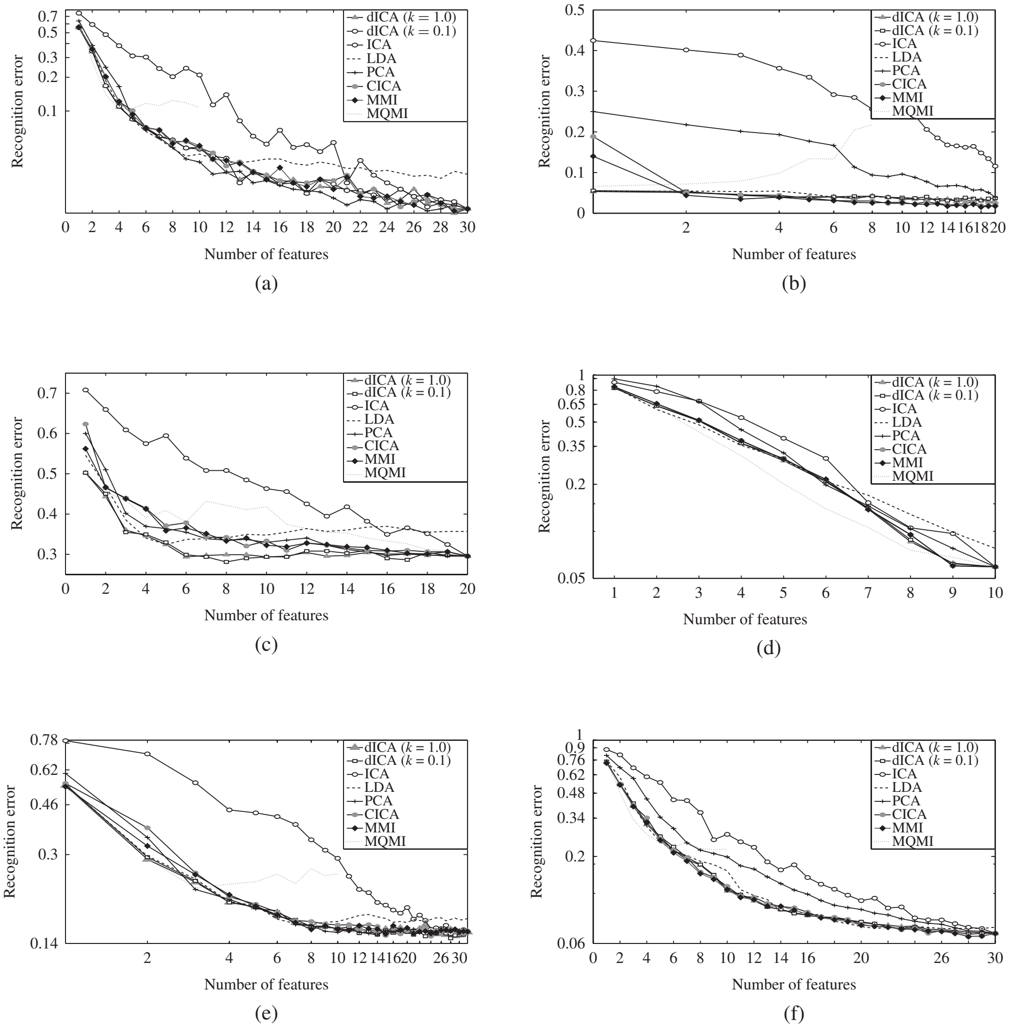
To compare the ability of the algorithm to correctly associate data samples to the classes, the scatter plot of 2-D dICA features is shown along with the 2-D scatter plots of LDA, PCA, and ICA. The learning rate η in (9) and (14) is chosen as 0.1.

From Fig. 1, it is seen that dICA and LDA features show better clustering performance in comparison to low-dimensional linear feature space obtained using unsupervised PCA and ICA. The mean classification performance is obtained over four random experiments with 108 test samples of male and female face images. With 2-D PCA and ICA features, we achieve a classification performance of 78.24% and 59.84%, respectively. On the other hand, 2-D dICA feature slightly outperforms LDA with a recognition rate of 94.9% over 94.6% and gives much better performance over PCA and ICA features.

TABLE II

MAXIMUM RECOGNITION PERFORMANCE IN %. THE VALUES IN THE PARANTHESIS (·) ARE THE NUMBER OF FEATURES REQUIRED TO OBTAIN THE CORRESPONDING RECOGNITION PERFORMANCE. SVM WITH GAUSSIAN KERNAL IS USED AS CLASSIFIER. GENDER RECOGNITION WITH 1-D CONVENTIONAL LDA FEATURE IS OBTAINED USING EUCLIDEAN DISTANCE MEASURE WHEN 1-NN CLASSIFIER IS USED

Database	L	Classifier	*d*ICA $k = 1$	*d*ICA $k = 0.1$	LDA	PCA	ICA	Conventional LDA
UCI Handwritten digits	30	SVM	**98.78** (29)	98.66 (30)	97.50 (28)	98.72 (27)	98.66 (30)	96.05 (9)
		1-NN	**97.22** (28)	97.05 (26)	95.44 (22)	**97.22** (23)	96.99 (27)	94.60 (9)
German emotion speech (AIBO features)	20	SVM	70.63 (6)	**71.89** (8)	67.57 (5)	70.45 (19)	70.45 (20)	66.31 (6)
		1-NN	60.00 (7)	**60.18** (7)	58.56 (6)	56.04 (7)	51.53 (20)	58.56 (6)
Postech faces gender recognition	50	SVM	**97.22** (20)	96.99 (13)	96.88 (19)	96.06 (20)	88.43 (20)	94.68 (1)
		1-NN	**93.63** (11)	93.52 (5)	93.29 (11)	90.74 (20)	83.45 (19)	90.63 (1)
Letter recognition	10	SVM	**94.07** (10)	**94.07** (10)	92.23 (10)	**94.07** (10)	**94.07** (10)	92.23 (10)
		1-NN	**92.37** (9)	92.07 (10)	**92.37** (10)	89.62 (10)	92.35 (9)	**92.37** (10)
Multiple feature dataset	30	SVM	85.05 (27)	**85.35** (26)	83.55 (8)	84.95 (14)	84.75 (26)	83.20 (9)
		1-NN	**81.65** (26)	81.30 (11)	80.00 (28)	81.45 (9)	81.50 (29)	79.10 (9)
Isolated spoken letter recognition	30	SVM	93.10 (30)	**93.18** (28)	92.62 (27)	93.10 (30)	93.10 (30)	92.51 (25)
		1-NN	86.34 (26)	**86.43** (24)	86.13 (26)	85.78 (30)	85.78 (30)	85.68 (25)



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 2. Comparing recognition performance using different number of features. (a) Optical handwritten digits. (b) Postech gender recognition from faces. (c) German emotion speech dataset. (d) Written letter (alphabet) recognition. (e) Multiple feature for numerical. (f) Isolated spoken alphabet datasets.

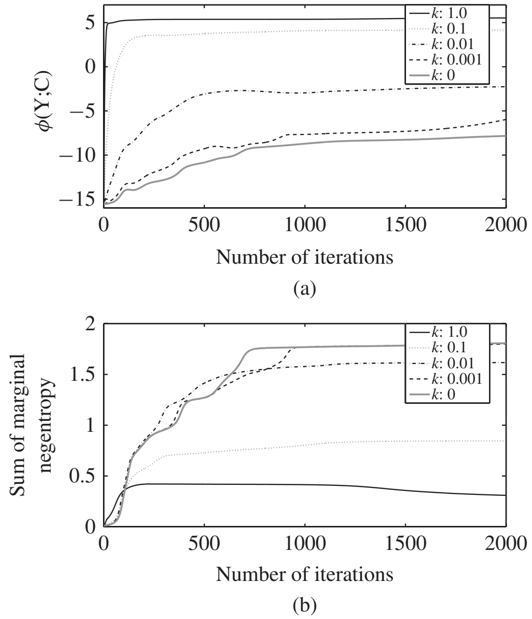
Table II shows the maximum recognition performance in % obtained by SVM and 1-NN classifier with cosine similarity. The value in the paranthesis $(\cdot)$ is the number of features that obtains the corresponding maximum recognition. It is seen that maximum performance is obtained when *d*ICA features are used for classification in comparison to other linear feature extraction methods. Experimental results under LDA section are obtained by using the subspace search method proposed by Murillo *et al.* [32]. In the case of conventional LDA, orthonormal features are the eigenvectors of $\mathbf{S}_W^{-1}\mathbf{S}_B$. The $(C-1)$ eigenvectors maximize the determinant of $\mathbf{S}_W^{-1}\mathbf{S}_B$. It is worth noting that LDA has the tendency to overfit to training data and may not necessarily be good for data representation. In case of the letter recognition dataset, the cardinality of LDA feature subspace is maximally confined to $L = 10$. The original data dimension is $P = 16$, which is later reduced to 10 using PCA. Lack of sufficient attributes prevents us from extracting all the meaningful features for letter recognition dataset.

Fig. 2 compares the recognition performance of *d*ICA with LDA, ICA based on Negentropy maximization, and PCA. Classification performance with different numbers of extracted features is shown when SVM is used as the classifier and the maximum recognition performance in the $\chi - \sigma$ parameter space of SVM classifier is used to compare the performance. It is clearly seen that *d*ICA outperforms PCA and ICA for any number of extracted features, $R < L$. Although, recognition performance of *d*ICA is comparable with that of LDA, it largely gives improved recognition performance for different numbers of extracted features. Recognition performance using the LDA feature space with cardinality around $(C-1)$ shows slight improvement over the same number of *d*ICA features. However, as shown in Fig. 2, the performance improvements of LDA features are much less when compared to the performance improvements achieved using more *d*ICA features.

*d*ICA show improved or at least similar recognition performance in comparison to CICA and Murillo's MMI approach as can be seen from Fig. 2. The performance of features extracted by maximization of QMI degrades as the feature dimension increases. This degradation in recognition performance can be accounted by the use of multivariate Gaussian kernels for estimation of probability distribution. Especially, in case of emotion recognition from the German speech database, the performance of MQMI features is the worst because of the lack of sufficient samples to model the distribution of data using multivariate Gaussian kernel.

V. Discussion

The experimental results show improved recognition performance when *d*ICA features are used for classification. LDA can theoretically extract only $(C-1)$ features. Searching for the LDA features in subspace with cardinality greater than $(C-1)$ is performed by finding orthonormal features in the complementary space using Gram–Schmidt orthonormalization. This procedure results in overfitting of the LDA feature space recursively in accordance with the training data without additional discriminant properties. Therefore, not much improvement is seen in recognition performance as the feature space increases over $(C-1)$.



(a)

(b)

Fig. 3. Convergence curve for *d*ICA with $R = C - 1 = 9$ for optical handwritten digits dataset. (a) Discriminant function, $\phi(\mathbf{Y}; C)$. (b) Marginal Negentropy sum, $\sum_i^R J(\mathbf{y}_i)$.

Maximization of the objective function mentioned in (8) finds the stationary point solution that maximizes the non-Gaussianity of the extracted latent variables (features). However, these extracted features do not ensure good classification performance as they do not consider class discrimination. On the other hand, *d*ICA features are obtained by the maximization of (13). While maximization of the sum of marginal Negentropy finds independent latent variables with good data generalization, maximization of FLD finds a good discriminant model. Fig. 3 compares the convergence curves of the summation of marginal Negentropy and the discriminant function $\phi(\mathbf{W}, \mathbf{Z}, C)$ of *d*ICA and ICA features with $R = C - 1 = 9$. The features are extracted from the optical handwritten digit dataset.

The stationary point solution of *d*ICA finds a feature space with cardinality R whose sum of marginal Negentropy is less than those extracted using ICA learning in (9). It can be seen in Fig. 3. However, *d*ICA features show much better class discrimination compared to ICA features. Although *d*ICA does not find globally independent features, it is still able to remove higher order statistical dependences among the features.

Looking back at the objective function of *d*ICA given in (13), both ICA and LDA are its special cases. Setting the value of $k = 0$ gives the objective function of ICA. On the other hand, setting a high value of k ignores the Negentropy maximization term and finds the LDA feature. From Fig. 3 it can be seen that choosing a small value of k will extract latent variables with reduced redundancy. However, a lower value of k may not ensure good discriminative features. The convergence of the algorithm can be further improved by using an adaptive learning rate η. The computational complexity of

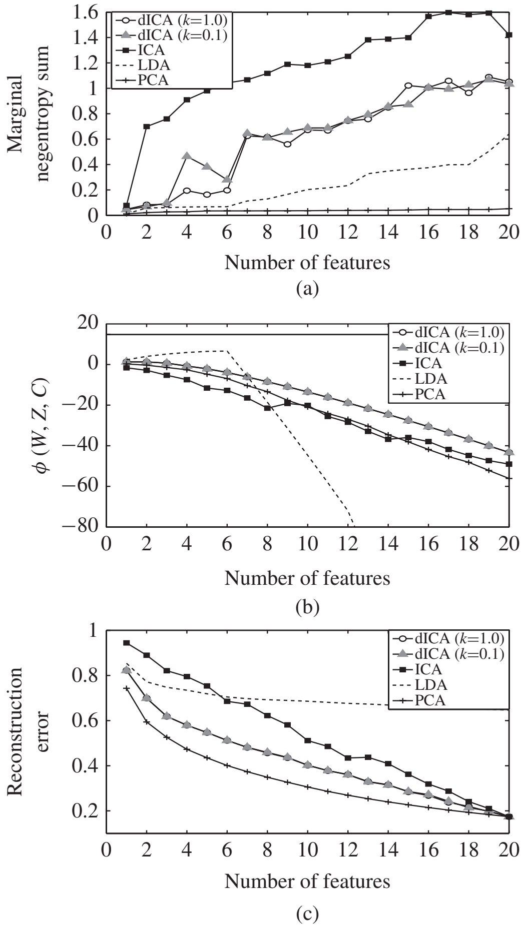


Fig. 4. (a) Marginal Negentropy sum, $\sum_i^R J(\mathbf{y}_i)$. (b) Discriminant function, $\phi(\mathbf{W}, \mathbf{Z}, C)$. (c) Data reconstruction error for German emotion speech dataset. (d) Statistical significance of *d*ICA with other feature extraction methods. Recognition experiments are performed on cross-validation sets obtained from the German emotion speech database.

ICA and *d*ICA update rule given in (9) and (14) is O(RLN) for each iteration.

The extracted *d*ICA features show much better class discrimination compared to ICA features, as *d*ICA jointly maximizes the discriminant function defined in (20). Fig. 4(a) shows the sum of marginal Negentropy at convergence for *d*ICA, ICA, LDA, and PCA features extracted from the German emotion speech database. It is seen that maximum non-Gaussianity was obtained using ICA features, which is a measure of independence among the features. Although *d*ICA features with $k = \{1, 0.1\}$ does not find globally independent features, it is still able to remove higher order statistical dependencies among the features.

In Fig. 4(b), the discriminant function, i.e., the sum of logarithm of marginal FLD, is compared for *d*ICA, ICA, LDA, and PCA features extracted from the German emotion speech database. Both *d*ICA and ICA learning algorithms have the same initialization $\mathbf{W}_0$ (random initialization (*randn* function in MATLAB with *seed* = 1)). It is seen that the optimal value of $\phi(\mathbf{W}^{opt}, \mathbf{Z}, C)$ at convergence, i.e., $\mathbf{W} = \mathbf{W}^{opt}$, is higher for *d*ICA features in comparison to ICA features. Also, the converged value shows improved discriminative power of features than the orthonormal PCA and LDA features.

As can be seen from Fig. 4(c), data reconstruction error for the German emotion speech database with *d*ICA is less compared to the linear decomposition of data using LDA. ICA features are optimized to maximize the non-Gaussianity of the features and do not consider data reconstruction error. On the other hand, PCA features are optimized to minimize the data reconstruction error. However, both ICA and PCA do not give a good discriminant model as can be seen from Figs. 1 and 2. *d*ICA features show not only good classification performance compared to other approaches (PCA, ICA, LDA) but also improved data reconstruction performance.

The *d*ICA approach takes the benefit of both LDA and ICA approaches. As the cardinality of LDA feature space increases above $(C - 1)$, there is not much improvement in discrimination ability of features, and the deflationary approach using Gram–Schmidt orthonormalization does not improve the representative ability of features to reproduce data with minimal reconstruction error. From Fig. 4, it appears that the low-dimension feature space of *d*ICA is mostly dominated by the discriminant function in the learning rule. However, as the cardinality of feature space goes above $C - 1$, the Negentropy maximization term of the objective function dominates, extracting features with maximum non-Gaussianity while preserving the discriminative ability of model. Similar trends are also observed for other datasets considered in this paper.

A. Comparison with CICA, MMI, and MQMI

Both CICA and MMI can be considered as special cases of *d*ICA when MI is used to define the discriminant function $\phi(\mathbf{W}, \mathbf{Z}, C)$, which is given as

$$\phi(\mathbf{W}, \mathbf{Z}, C) = I(\mathbf{Y}; C). \tag{29}$$

Under the independence assumption of extracted *d*ICA features, joint distribution of multivariate feature $\mathbf{Y}$ can be factorized as $\prod_i^R p(\mathbf{y}_i)$, which results in the simplified formulation of the functional measure of classification performance as

$$\phi(\mathbf{W}, \mathbf{Z}, C) = \sum_{i=1}^{R} \sum_{c=1}^{C} I(\mathbf{y}_i; c). \tag{30}$$

The objective function using the discriminant function in (30) is given as

$$\begin{aligned} \hat{J}(\mathbf{Y}) \;=\; & (1-k)\sum_{i=1}^{R} J(\mathbf{y}_i) + kR\alpha + k\sum_{i}^{R}\sum_{c=1}^{C} p(c) J(\mathbf{y}_i|c) \\ & -k\sum_{i}^{R}\sum_{c=1}^{C} p(c)\left(\frac{1}{2}\log(2\pi\sigma_{ic}^2)\right) \end{aligned} \tag{31}$$

where α is a constant which is the entropy of univariate Gaussian random variable. Equation (31) shows that the maximization of marginal sum of Negentropy of $\mathbf{y}_i$ will results in minimization of $\sum_i^R I(\mathbf{y}_i; C)$.

The choice of objective function as given in (31) would result in inferior recognition performance, as Negentropy

TABLE III
COMPARING MAXIMUM RECOGNITION PERFORMANCE OF *d*CIA, CICA [33], MAXIMIZATION OF MUTUAL INFORMATION [32], AND MAXIMIZATION OF QMI [31]

Database	*d*ICA	CICA	MMI	MQMI
UCI Handwritten	**98.78** (29)	98.66 (29)	98.72 (29)	89.54 (4)
German emotion speech	**71.89** (8)	70.45 (20)	70.45 (20)	70.45 (20)
Postech face gender	97.22 (20)	98.15 (18)	**98.38** (14)	93.40 (1)
Letter recognition	**94.07** (10)	**94.07** (10)	**94.07** (10)	**94.07** (10)
Multiple feature dataset	**85.35** (26)	85.10 (22)	84.80 (15)	77.50 (3)
Isolated spoken letter	93.18 (28)	93.38 (28)	**93.41** (28)	77.95 (9)

maximization would essentially result in minimization of (30). When $k = 1$, the objective function reduces to that of CICA. Maximization of (30) can be considered a special case of (31) when $k \to \infty$. MMI proposed by Murillo *et al.* validly considered that the residual term $[I(\mathbf{Y}) - I(\mathbf{Y}|C)]$ is negligible for PCA features that are used as input to the network. However, maximization of the MI criterion in (30) is achieved by the minimization of the marginal sum of Negentropy, which in turn maximizes $I(\mathbf{Y})$, and maximization of the marginal sum of Negentropy given a class that minimizes $I(\mathbf{Y}|C)$. Therefore, the assumption that residual term can be neglected is not valid and results in erroneous measure of $I(\mathbf{Y}; C)$.

In case of maximization of QMI, Gaussian kernels are used to estimate the conditional multivariate distribution of the extracted features. Unfortunately, as the cardinality of the feature space increases, the estimation becomes erroneous because of the curse of dimensionality as can be seen from Fig. 2. This performance degradation is due to the lack of sufficient samples per class to estimate the multivariate conditional density. In addition, the interaction among the samples reduces with increase in the cardinality of feature space because the volume of the hypersphere defined by the width of Gaussian kernels exponentially goes to 0.

Table III compares the maximum recognition performance of *d*ICA (better performance between $k = 1.0$ and $k = 0.1$ is reported) with CICA, MMI, and MQMI based feature extraction methods. The recognition performance is given in % and is obtained using SVM on the extracted features. Fig. 5 shows the p-value obtained by performing t-test on recognition results from different cross-validation experiments on the German emotion speech dataset. *d*ICA is compared with other feature extraction methods discussed in this paper. It is seen that the *d*ICA results are statistically significant for most of the feature dimensions. In particular, *d*ICA shows better statistical significance in comparison with ICA, LDA features with cardinality greater than $(C-1)$, and maximization of QMI. Since, the objective function of *d*ICA includes maximization of FLD, both *d*ICA and LDA have similar performance when the number of extracted discriminant features is close to $(C - 1)$. All the feature extraction methods show similar performance when the number of extracted features is large because of good data representation. Similar performance trends were seen for the other datasets.

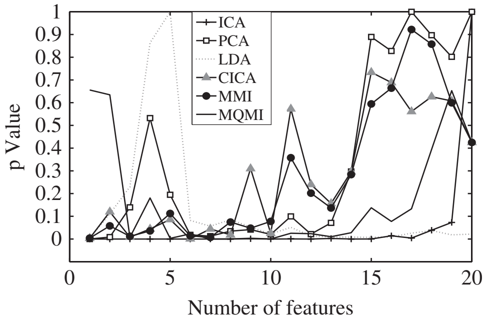


Fig. 5. Statistical significance of *d*ICA with other feature extraction methods. Recognition experiments are performed on cross-validation sets obtained from the German emotion speech database.

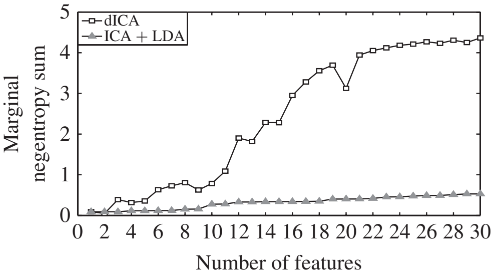


Fig. 6. Comparing Marginal Negentropy sum, $\sum_i^R J(\mathbf{y}_i)$ for Multiple Feature dataset.

B. Comparison with DICA Subspace

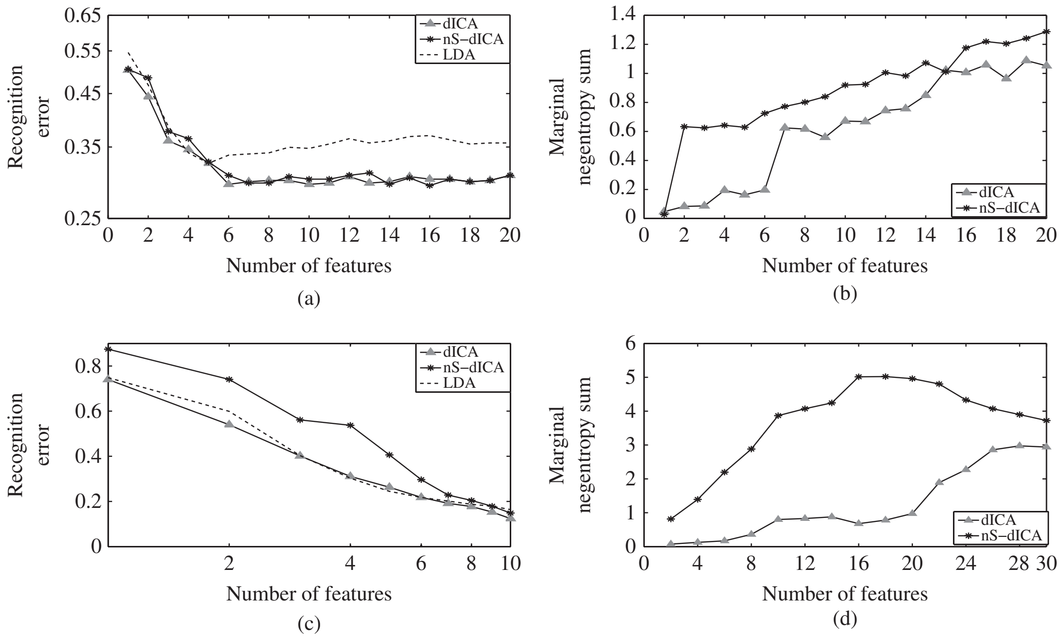
Fei *et al.* [27] performed ICA at first stage to extract L features and later projected the independent components to a discriminant space using LDA. Extraction of L ICA features at the first stage requires a computational complexity of O(L^{2}N) for each iteration. On the other hand, *d*ICA proposed in this paper requires O(RLN) computation for each iteration with $R \leq L$. *d*ICA proposed in this paper shows higher recognition performance as can be seen from Table IV.

The two-stage DICA approach makes the extracted features redundant, as the linear transformation by LDA will not preserve the statistical independence among the features. It can be seen from Fig. 6. *d*ICA proposed in this paper finds discriminant features with minimum redundancy, hence the name *d*ICA is more appropriate.

VI. NONSINGULAR *d*ICA (NS-*d*ICA) AND OPTIMAL *k*

Under the unit covariance constraint of statistically independent $\mathbf{Y}$ which also results in orthonormal $\mathbf{W}$, the sum of between-class and within-class scatter of feature matrix is an identity matrix for whitened data

$$\mathbf{W}^T \mathbf{S}_T \mathbf{W} = \mathbf{I}. \tag{32}$$



(a)

(b)

(c)

(d)

Fig. 7. Comparing recognition performance and marginal sum of Negentropy using nS-*d*ICA features with optimal k for German emotion speech and ISOLET dataset. (a) Recognition error for German emotion speech dataset. (b) Marginal Negentropy sum, $\sum_i^R J(\mathbf{y}_i)$ for German emotion speech dataset. (c) Recognition error for Isolated spoken alphabet dataset. (d) Marginal Negentropy sum, $\sum_i^R J(\mathbf{y}_i)$ for Isolated spoken alphabet dataset.

TABLE IV

COMPARING MAXIMUM RECOGNITION PERFORMANCE OF *d*ICA WITH DISCRIMINANT ICA BASED SUBSPACE METHOD PROPOSED BY FEI *et al.* [27]. THE VALUE OF k TO EXTRACT *d*ICA FEATURES IS 1. SVM IS USED AS THE CLASSIFIER

Database	UCI Handwritten Digits	German speech	Multiple Feature	ISOLET
*d*ICA	**98.66** (30)	**71.89** (8)	**85.35** (26)	**93.18** (28)
Fei *et al.* [27]	95.33	70.27	83.80	92.56

This results in modification of the discriminant function $F(\mathbf{W})$ to

$$F(\mathbf{W}) = \frac{\left|\mathbf{W}^T\mathbf{S}_B\mathbf{W}\right|}{\left|\mathbf{I}-\mathbf{W}^T\mathbf{S}_B\mathbf{W}\right|}. \tag{33}$$

Using the additive property of statistically independent features and orthonormality constraint, the functional measure of classification, $\phi(\mathbf{W},\mathbf{Z},C)$, can be given as [22]

$$\phi(\mathbf{W},\mathbf{Z},C) = \sum_{i=1}^{R}\phi(\mathbf{w}_i,\mathbf{Z},C) = \sum_{i=1}^{R}\log\frac{\mathbf{w}_i^T\mathbf{S}_B\mathbf{w}_i}{1-\mathbf{w}_i^T\mathbf{S}_B\mathbf{w}_i}. \tag{34}$$

Since (34) is dependent only on the value of $\mathbf{w}_i^T\mathbf{S}_B\mathbf{w}_i$, the optimal linear transformation which attempts to find maximally discriminant orthonormal features for whitened observations $\mathbf{Z}$ can be given as

$$\tilde{W} = \arg\max_{\mathbf{W}}\phi(\mathbf{W},\mathbf{Z},C) = \arg\max_{\mathbf{W}}\sum_{i=1}^{R}\log\mathbf{w}_i^T\mathbf{S}_B\mathbf{w}_i. \tag{35}$$

The partial derivative of modified $\phi(\mathbf{W},\mathbf{Z},C)$ in (35) with respect to basis vector $\mathbf{w}_{li}$ is given as

$$\frac{\partial\phi(\mathbf{W},\mathbf{Z},C)}{\partial w_{li}} = 2\sum_{c=1}^{C}\sum_{n\in Class\ c}\frac{(\mu_{ic}-\mu_i)}{\sum_{c'=1}^{C}N_{c'}\left(\mu_{ic'}-\mu_i\right)^2}z_{ln}. \tag{36}$$

Substituting the value of $\partial\phi(\mathbf{W},\mathbf{Z},C)/\partial\mathbf{w}_{li}$ derived in (36) back in (14) results in a discriminative feature extraction algorithm that maximizes the sum of marginal Negentropy along with the between-class scatter for each feature. Since we perform symmetric orthonormalization of $\mathbf{W}$, substituting β_i as zero in (14) does not affect the results at convergence. Nonsingular *d*ICA (nS-*d*ICA) attempts to remove redundancy among the projected features and improves the discrimination ability by maximally separating the class mean of transformed data on each projection. nS-*d*ICA also has reduced computational complexity in comparison to (23).

A. Optimal k for Nonsingular dICA

The Negentropy maximization objective function using (35) can be given as

$$\hat{J}(\mathbf{Y}) = \sum_{i=1}^{R}J\left(\mathbf{y}_i\right) + \sum_{i}^{R}k_i\log\mathbf{w}_i^T\mathbf{S}_B\mathbf{w}_i. \tag{37}$$

Taking the gradient of (37) with respect to a $\mathbf{w}_i$ and equating it to zero results in

$$2\left(\gamma_i E\left(\mathbf{Z}g\left(\mathbf{w}_i^T\mathbf{Z}\right)\right)\right) + 2k_i\frac{\mathbf{S}_B\mathbf{w}_i}{\mathbf{w}_i^T\mathbf{S}_B\mathbf{w}_i} = 0. \tag{38}$$

Multiplying $\mathbf{w}_i^T$ on both sides gives

$$k_i = -\gamma_i E\left(\mathbf{y}_i g\left(\mathbf{w}_i^T\mathbf{y}_i\right)\right). \tag{39}$$

Fig. 7(a) shows that no significant differences are seen in the recognition performance for the German emotion speech database using AIBO features. nS-*d*ICA with optimal k equally emphasizes the gradient of Negentropy and the discriminant function. It gives improved independence among features in comparison to the fixed values of k while maintaining the recognition performance, as can be seen from Fig. 7(b) for the German emotion speech dataset. While, the value of adaptive k given in (39) is an optimal value that maximizes the objective function in (37), it does not ensure that the best recognition performance will be obtained from it. As seen from Fig. 7(c), optimal k does not give good recognition performance for most of the low-dimensional *d*ICA features in the case of ISOLET database. A high recognition error is related to a lower value of discriminant function, which means that the optimal k would give large value of marginal Negentropy sum, which can be seen from Fig. 7(d) for ISOLET database.

In most of the experiments, it is seen that we do not need to overemphasize the discriminant part of learning, i.e., the value of k in between 0.1 and 1 would still give good performance. Batitti *et al.* had also found the weighting factor in a similar range [14]. It would be an interesting future research subject to find an adaptive k that sufficiently biases maximization of discriminant function while minimizing the redundancy among the features.

VII. Conclusion

*d*ICA is an adaptive framework of extraction of discriminative independent features that utilizes information of class variables. A gradient-based joint maximization of marginal sum of Negentropy and Fisher discriminant score was derived that inherently takes into consideration the orthonormality of the demixing vector $\mathbf{w}_i$. Recognition tasks with *d*ICA features result in improved classification performance over unsupervised PCA and ICA features. It also shows better recognition performance in comparison to supervised methods like LDA and information theoretic approaches that maximize MI between multivariate feature and class. *d*ICA features not only minimize redundancy in the feature space but also give a better representative model with less data reconstruction error. Reduced entropy of *d*ICA features can be used for efficient coding.

Appendix

Derivation of update rule for maximization of the discriminant function $\phi(\mathbf{W}, \mathbf{Z}, C)$ given in (21) with respect to $\mathbf{W}$

$$\phi(\mathbf{W}, \mathbf{Z}, C) = \sum_{i}^{R} \log \frac{\mathbf{w}_i^T \mathbf{S}_B \mathbf{w}_i}{\mathbf{w}_i^T \mathbf{S}_W \mathbf{w}_i} \tag{40}$$

$$\begin{aligned}\phi(\mathbf{W}, \mathbf{Z}, C) = &\sum_{i=1}^{R} \log \sum_{c=1}^{C} N_c \left(\mu_{ic} - \mu_i\right)^2 \\ &- \sum_{i=1}^{R} \log \sum_{c=1}^{C} N_c \sigma_{ic}^2,\end{aligned} \tag{41}$$

$$\phi(\mathbf{W}, \mathbf{Z}, C) = \pi - \varpi \tag{42}$$

$$\frac{\partial \phi(\mathbf{W}, \mathbf{Z}, C)}{\partial w_{li}} = \frac{\partial \pi}{\partial w_{li}} - \frac{\partial \varpi}{\partial w_{li}} \tag{43}$$

where w_{li} is the lth element of column vector $\mathbf{w}_i$.

Gradient of π with respect to $\mathbf{w}_i$ $[\nabla \pi]$:

Using the chain rule, the partial derivation of π with respect to w_{li} can be given as

$$\frac{\partial \pi}{\partial w_{li}} = \sum_{n=1}^{N} \left(\sum_{c=1}^{C} \frac{\partial \pi}{\partial \mu_{ic}} \frac{\partial \mu_{ic}}{\partial y_{in}} \right) \frac{\partial y_{in}}{\partial w_{li}}. \tag{44}$$

$$\frac{\partial \pi}{\partial \mu_{ic}} = \frac{2N_c \left(\mu_{ic} - \mu_i\right)}{\sum_{c'=1}^{C} N_{c'} \left(\mu_{rc'} - \mu_i\right)^2} \tag{45}$$

$$\frac{\partial \mu_{ic}}{\partial y_{in}} = \frac{1}{N_c} \quad \text{if } n \in \text{ Class c}, \ 0 \ \text{ otherwise} \tag{46}$$

$$\frac{\partial y_{in}}{\partial w_{li}} = z_{ln}. \tag{47}$$

Substituting (45)–(47) in (44) gives

$$\frac{\partial \pi}{\partial w_{li}} = \sum_{c=1}^{C} \sum_{n \in Class\ c} \frac{2\left(\mu_{ic} - \mu_i\right)}{\sum_{c'=1}^{C} N_{c'} \left(\mu_{ic'} - \mu_i\right)^2} z_{ln}. \tag{48}$$

Gradient of ϖ with respect to $\mathbf{w}_i$ $[\nabla \varpi]$:

Using the chain rule, the partial derivation of ω with respect to w_{li} can be given as

$$\frac{\partial \varpi}{\partial w_{li}} = \sum_{n=1}^{N} \left(\frac{\partial \varpi}{\partial y_{in}} \right) \frac{\partial y_{in}}{\partial w_{li}} \tag{49}$$

$$\frac{\partial \varpi}{\partial y_{in}} = \frac{\frac{\partial}{\partial y_{in}} \left(\sum_{n'=1}^{N} \left(y_{in'} - \mu_{ic(n')} \right)^2 \right)}{\sum_{n'=1}^{N} \left(y_{in'} - \mu_{ic(n')} \right)^2} \tag{50}$$

where $c(n')$ is the class index of sample n'.

Let $\alpha = \sum_{n'=1}^{N} \left(y_{in'} - \mu_{ic(n')} \right)^2$, then (50) can be written as

$$\frac{\partial \varpi}{\partial y_{in}} = 2\alpha^{-1} \left(y_{in} - \mu_{ic(n)} \right). \tag{51}$$

Using (51) and (47) and substituting them in (49) give the partial derivative of ω with respect to w_{li}

$$\frac{\partial \varpi}{\partial w_{li}} = \sum_{n=1}^{N} \frac{2\left(y_{in} - \mu_{ic(n)} \right)}{\sum_{n'=1}^{N} \left(y_{in'} - \mu_{ic(n')} \right)^2} z_{ln} \tag{52}$$

$$\frac{\partial \varpi}{\partial w_{li}} = \sum_{c=1}^{C} \sum_{n \in Class\ c} \frac{2\left(y_{in} - \mu_{ic} \right)}{\sum_{c'=1}^{C} \sum_{n' \in Class\ c'} \left(y_{in'} - \mu_{ic'} \right)^2} z_{ln}. \tag{53}$$

Substituting the values of (53) and (48) in (43) gives the partial derivative of $\phi(\mathbf{W}, \mathbf{Z}, C)$ with respect to w_{li}.

References

[1] G. E. Hinton and T. J. Sejnowski, *Unsupervised Learning: Foundations of Neural Computation*. Cambridge, MA: MIT Press, 1999.

[2] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4–37, Jan. 2000.

[3] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.

[4] S. Haykins, *Unsupervised Adaptive Filtering*. New York: Wiley, 2000.

[5] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognit. Neurosci.*, vol. 3. no. 1, pp. 71–86, 1991.

[6] M. Kirby and L. Sirovich, "Application of the Karhunen–Loève procedure for the characterization of human faces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 1, pp. 103–108, Jan. 1990.

[7] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001.

[8] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, no. 6, pp. 1129–1159, Nov. 1995.

[9] T. W. Lee, *Independent Component Analysis: Theory and Applications*. Boston, MA: Kluwer, 1998.

[10] P. Comon, "Independent component analysis, a new concept?" *Signal Process.*, vol. 36, no. 3, pp. 287–314, Apr. 1994.

[11] S. Roberts and R. Everson, *Independent Component Analysis: Principles and Practice*. Cambridge, U.K.: Cambridge Univ. Press, 2001.

[12] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, nos. 1–2, pp. 273–324, Dec. 1997.

[13] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski, "Face recognition by independent component analysis," *IEEE Trans. Neural Netw.*, vol. 13, no. 6, pp. 1450–1464, Nov. 2002.

[14] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Trans. Neural Netw.*, vol. 5, no. 4, pp. 537–550, Jul. 1994.

[15] N. Kwak and C.-H. Choi, "Input feature selection by mutual information based on Parzen window," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 12, pp. 1667–1671, Dec. 2002.

[16] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[17] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *J. Mach. Learn. Res.*, vol. 5, pp. 1205–1224, Oct. 2004.

[18] D.-S. Huang and C.-H. Zheng, "Independent component analysis-based penalized discriminant method for tumor classification using gene expression data," *Bioinformatics*, vol. 22, no. 15, pp. 1855–1862, May 2006.

[19] C. S. Dhir and S. Y. Lee, "Hybrid feature selection: Combining fisher criterion and mutual information for efficient feature selection," in *Proc. Adv. Neural Inf. Process.*, vol. 5506. 2009, pp. 613–620.

[20] A. M. Martinez and A. C. Kak, "PCA versus LDA," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 228–233, Feb. 2001.

[21] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Face recognition using LDA-based algorithms," *IEEE Trans. Neural Netw.*, vol. 14, no. 1, pp. 195–200, Jan. 2003.

[22] K. Fukunaga, *Introduction of Statistical Pattern Recognition*, 2nd ed. New York: Academic, 1990.

[23] K.-C. Li, "Sliced inverse regression for dimension reduction," *J. Amer. Stat. Assoc.*, vol. 86, no. 414, pp. 316–327, Jun. 1991.

[24] D. R. Hardoon, S. Szedmak, and J. R. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, Dec. 2004.

[25] J. D. R. Farquhar, D. R. Hardoon, H. Meng, J. Shawe-Taylor, and S. Szedmak, "Two view learning: SVM-2K, theory and practice," in *Proc. Adv. Neural Inf. Process. Syst. 18*, 2006, pp. 355–362.

[26] P. N. Belhumeur, J. P. Hespanda, and D. J. Kriegeman, "Eigenfaces versus fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[27] F. Long, J. He, X. Ye, Z. Zhuang, and B. Li, "Discriminant independent component analysis as a subspace representation," *J. Electron. (China)*, vol. 23, no. 1, pp. 103–106, 2006.

[28] A. Renyi, *Probability Theory*. Amsterdam, The Netherlands: North Holland, 1970.

[29] K. Torkkola, "Feature extraction by non parametric mutual information maximization," *J. Mach. Learn. Res.*, vol. 3, pp. 1415–1438, Mar. 2003.

[30] J. Principe, D. Xu, and J. Fisher, "Information theoretic learning," in *Unsupervised Adaptive Filtering*, vol. 1. New York: Wiley, 2000.

[31] K. E. Hild, D. Erdogmus, K. Torkkola, and J. C. Principe, "Feature extraction using information-theoretic learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 9, pp. 1385–1392, Sep. 2006.

[32] J. M. Leiva-Murillo and A. Artes-Rodriguez, "Maximization of mutual information for supervised linear feature extraction," *IEEE Trans. Neural Netw.*, vol. 18, no. 5, pp. 1433–1441, Sep. 2007.

[33] S. Akaho, "Conditionally independent component analysis for supervised feature extraction," *Neurocomputing*, vol. 49, nos. 1–4, pp. 139–155, Dec. 2002.

[34] N. Slonim and N. Tishby, "Agglomerative information bottleneck," in *Proc. Adv. Neural Process. Syst. 12*, 2000, pp. 617–623.

[35] U. Amato, A. Antoniadis, and G. Gregoire, "Independent component discriminant analysis," *Int. J. Math.*, vol. 3, no. 7, pp. 735–753, 2003.

[36] M. Dyrholm, C. Christoforou, and L. C. Parra, "Bilinear discriminant component analysis," *J. Mach. Learn. Res.*, vol. 8, pp. 1097–1111, May 2007.

[37] E. K. P. Chong and S. H. Zak, *An Introduction to Optimization*, 2nd ed. New York: Wiley, 2001.

[38] A. Asuncion and D. J. Newman. (2007). *UCI Machine Learning Repository*. School Inf. Comput. Sci., Univ. California, Irvine [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html

[39] S. Aeberhard, D. Coomans, and O. de Vel, "Comparison of classifiers in high dimensional settings," Dept. Comput. Sci. & Math. Stat., James Cook Univ. North Queensland, Townsville, Australia, 1992.

[40] E. Alpaydin and C. Kaynak, "Cascaded classifiers," *Kybernetika*, vol. 34, no. 4, pp. 369–374, 1998.

[41] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Proc. Interspeech*, Sep. 2005, pp. 1517–1520.

[42] O. Pierre-Yves, "The production and recognition of emotions in speech: Features and algorithms," *Int. J. Human-Comput. Stud.*, vol. 59, nos. 1–2, pp. 157–183, Jul. 2003.

[43] H.-S. Lee, S. Park, B.-N. Kang, J. Shin, J.-Y. Lee, H. Je, B. Jun, and D. Kim, "The POSTECH face database (PF07) and performance evaluation," in *Proc. IEEE Int. Conf. Autom. Face & Gesture Recognit.*, Sep. 2008, pp. 1–6.

[44] P. W. Frey and D. J. Slate, "Letter recognition using Holland-style adaptive classifiers," *Mach. Learn.*, vol. 6, no. 2, pp. 161–182, Mar. 1991.

[45] M. P. W. Breukelen, D. M. J. Tax, and J. E. Hartog, "Handwritten digit recognition by combined classifiers," *Kybernetika*, vol. 34, no. 4, pp. 381–386, 1998.

[46] M. A. Fanty and R. Cole, "Spoken letter recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 1990, pp. 220–226.

[47] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

[48] T. Joachims, "Making large-scale SVM learning practical," in *Advances in Kernel Methods-Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, Eds. Cambridge, MA: MIT Press, 1999.

[49] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood components analysis," in *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou, Eds. Cambridge, MA: MIT Press, 2005, pp. 513–520.

Chandra Shekhar Dhir received the B.Tech. degree in electrical engineering from the Indian Institute of Technology, Chennai, India, in 2002, and the M.S. and Ph.D. degrees in bio and brain engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2006 and 2010, respectively.

He was a Junior Manager in the Research and Development Section, Jindal Vijayanagar Steel Limited (now JSW), Karnataka, India, from July 2002 to April 2003, before joining the Brain Science Research Center, KAIST, as an Internship Trainee, from August 2003 to February 2004. Currently, he is with LG Electronics, Seoul, Korea, as a Senior Research Engineer. His current research interests include signal processing, information theory, blind source separation, supervised and unsupervised machine learning methods, and computer vision with a focus on 2-D object detection and recognition.

Soo-Young Lee (M'83) received the B.S. degree from Seoul National University, Seoul, Korea, in 1975, the M.S. degree from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 1977, and the Ph.D. degree from the Polytechnic Institute of New York, Brooklyn, in 1984.

He was with Taihan Engineering Company, Seoul, from 1977 to 1980. He was with the General Physics Corporation, Columbia, MD, from 1982 to 1985. In 1986, he joined the Department of Electrical Engineering, KAIST, as an Assistant Professor, and now is a Full Professor in the same department and also in the Department of Bio and Brain Engineering. From June 2008 to 2009, he was with the Mathematical Neuroscience Laboratory, RIKEN Brain Science Institute, Saitama, Japan, on sabbatical leave. His current research interests include artificial brains, human-like intelligent systems/robots based on biological information processing mechanisms in brains, mathematical models, neuromorphic chips, real-world applications, intelligent man-machine interfaces with electroencephalograms, and eye gaze.

Prof. Lee is a Past President of the Asia-Pacific Neural Network Assembly (APPNA) and has contributed to the International Conference on Neural Information Processing as a Conference Chair in 2000, the Conference Vice Co-Chair in 2003, and the Program Co-Chair in 1994 and 2002. He is on the Editorial Boards for the *Neural Processing Letters* and *Cognitive Neurodynamics* Journals. He received the Leadership Award and Presidential Award from the International Neural Network Society in 1994 and 2001, respectively, and the APPNA Service Award and Outstanding Achievement Award from APNNA in 2004 and 2009, respectively. From the International Society for Optical Engineering, he received the Biomedical Wellness Award and the Independent Component Analyses Unsupervised Learning Pioneer Award in 2008 and 2010, respectively.