INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 13(1), 24-38, February 2018.

Selective Feature Generation Method for Classification of Low-dimensional Data

S.-I. Choi, S.T. Choi, H. Yoo

Sang-Il Choi, Haanju Yoo
Department of Computer Science and Engineering
Dankook University
152, Jukjeon-ro, Suji-gu, Yongin-si
Gyeonggi-do, 16890, Korea
choisi@dankook.ac.kr, haanju.yoo@gmail.com

Sang Tae Choi*
Department of Internal Medicine
Chung-Ang University College of Medicine
102 Heukseok-ro, Dongjak-gu
Seoul, 06974, Korea.
*Corresponding author: beconst@cau.ac.kr

Abstract: We propose a method that generates input features to effectively classify low-dimensional data. To do this, we first generate high-order terms for the input features of the original low-dimensional data to form a candidate set of new input features. Then, the discrimination power of the candidate input features is quantitatively evaluated by calculating the 'discrimination distance' for each candidate feature. As a result, only candidates with a large amount of discriminative information are selected to create a new input feature vector, and the discriminant features that are to be used as input to the classifier are extracted from the new input feature vectors by using a subspace discriminant analysis. Experiments on low-dimensional data sets in the UCI machine learning repository and several kinds of low-resolution facial image data show that the proposed method improves the classification performance of low-dimensional data by generating features.
Keywords: feature generation, input feature selection, feature extraction, discriminant distance, low-dimensional data, data classification.

1 Introduction

Advances in information technology have resulted in a rapid increase in the amount of digital data that is available, and a significant amount of research has been carried out to develop tools to extract useful and necessary information from vast amounts of data. Such tools are currently being applied in various fields, including biometrics (e.g., iris, fingerprint and face recognition), data mining, diagnosis systems and pattern classification [22, 26].

When working with data samples, which are represented as 'input features', feature extraction methods can effectively improve classification performance by extracting useful information. When there are input features in a data sample, feature extraction methods find projection vectors to get new features containing the maximal information for problem solving [4, 14, 15, 27, 31]. Then, an input data sample is represented by a set of new features (feature vector), each of which is a linear combination of the input features.

The different feature extraction methods have different properties, and the appropriate method must be used corresponding to the characteristics of the data and the problem that is to be solved, e.g., data representation, classification, restoration, etc. Common feature extraction methods such as Principal Component Analysis (PCA) [27] and Linear Discriminant Analysis

(LDA) [15] have been the basis to develop other methods, including Null space LDA (NLDA) [4], Biased Discriminant Analysis (BDA) [31], etc. In these methods, data is stored in vector form, and the appropriate features are extracted using a covariance matrix which is appropriately defined depending on the problem to be solved. Methods such as MatFLDA [5], Two-Dimensional LDA (2DFLD) [29], Composite LDA (C-LDA) [18] and Composite BDA (C-BDA) [17] use an image covariance matrix instead of the covariance matrix. These image covariance-based methods can be used effectively for data in which input features are strongly correlated [17]. C-LDA can be viewed as a generalized image covariance-based method because C-LDA becomes identical to the 2DLDA or MatFLDA form when the composite vector is defined as a row or column vector.

In classification problems, an object is described as an array of attributes to search for the underlying patterns in the object. These attributes are represented as numerical values, which are stored in a vector form (input feature vector) [8]. For example, for blood test data for a person in a hospital, the dimension of the data is the number of test items. Even when using the same object, the attributes can be defined in different ways depending on the problem that is to be classified. For example, when classifying a dog, attributes such as food or skeletal structure can be used to classify species of mammals, amphibians, and the like, and when distinguishing individual objects belonging to the same group of animals, attributes such as hair color, size, age, etc. can be used. However, when expressing an object with attributes in this manner, the number of attributes is limited, and it is usually represented using low-dimensional data. On the other hand, temporal sensing data such as speech, or spatial data such as images is usually stored as high-dimensional data. Even such data is often reduced and stored as low-dimensional data such as a thumbnail image in order to effectively use the data in a small device, which has a relatively small computing power.

Most feature extraction methods mentioned above use a statistical correlation of input features and extract features from the shape information of the pixels constituting the image, so their classification performance is limited when the number of input features is too small and is affected by the resolution of the image. In the case of the DCV method, which offers a high performance for generic high-dimensional data, the dimension of the null space may decrease or disappear when the dimension of the data decreases. Therefore, it is necessary to generate meaningful features from the input features to effectively utilize the existing data classification techniques with low-dimensional data.

In this paper, we propose an input feature generation method for classification of low-dimensional data. According to the Theorem of Cover [10], if data samples are not distributed linearly and separably, they can be made into a linearly separated distribution through conversion into higher dimensions. Many methods use kernel functions to convert low-dimensional data into higher dimensions [9, 24, 28, 30]. These methods use a kernel matrix instead of directly computing kernel functions because doing so would require extensive computation. However, in this case, since the value of the high-dimensional data that is created can not be confirmed, even if the feature corresponding to the individual dimension of the high-dimensional data includes unnecessary information that do not help in classification, they can not be removed or separately used. In the proposed method, new input features are generated by adding a higher order term of individual input features, and the separability power for the original input features and the generated input features is measured using the discriminant distance scale [21]. Then, only features with high discrimination information are selectively used during data classification. The new input features improves the performance of existing discriminant feature extraction methods especially when classifying low-dimensional data. We recently investigated the feature generation method for face recognition and presented preliminary results in [6]. In this paper, we provide a more detailed analysis of the method, as well as an extensive discussion, and we apply the method to other classification problems other than face recognition. Through experiments on

various low-dimensional data sets, we confirmed that the classification performance is improved when using the proposed input feature generation method. The results of the experiment for low-resolution facial images show that the proposed method offers a higher recognition rate than when the resolution of such images is increased via interpolation.

This paper is organized as follows. In the next section, we examine the effect of the data dimension on the classification performance. Then, we describe the feature generation method and the optimal input feature selection method. Finally, the experimental results are described and the conclusion follows.

2 Effect of data dimensionality on classification performance

2.1 Subspace discriminant analysis

Subspace discriminant analysis methods represent a data sample as an n-dimensional vector $\mathbf{x}$. LDA, NLDA and BDA are representative methods of these subspace discriminant analysis methods. When there are N data samples with C classes and N_i samples for each class $c_i (i = 1, ..., C)$, the within class scatter matrix S_W and the between class scatter matrix S_B can be defined as follows:

$$\begin{aligned} S_W &= \sum_{i=1}^{C} \sum_{\mathbf{x_k} \in c_i} (\mathbf{x_k} - \boldsymbol{\mu}_i)(\mathbf{x}_k - \boldsymbol{\mu}_i)^T, \\ S_B &= \sum_{i=1}^{C} N_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T \end{aligned} \tag{1}$$

where $\boldsymbol{\mu}_i$ is the average of the samples in class c_i and $\boldsymbol{\mu}$ is the average of all N samples.

LDA finds a projection matrix $W_{Fisher} = [\boldsymbol{\mu}_1, \cdots, \mathbf{w}_{C-1}]$ consisting of projection vectors $\mathbf{w}_l (l = 1, \cdots, C-1)$ that satisfies the following objective function. This means that the LDA constructs a feature space that maximizes the covariance between the other classes while minimizing the covariance between the same classes in the range space of S_W [15]. W_{LDA} can be obtained by calculating the eigenvectors of $S_W^{-1} S_B$.

$$W_{Fisher} = argmax_W \frac{\left|W^T S_B W\right|}{|W^T S_W W|} \tag{2}$$

Unlike the LDA, which uses the range space of S_W, the NLDA uses the null space of S_W containing more discriminating information [4]. That is, a projection matrix W_{DCV} satisfying the following objective function is obtained in a space of $\left|W^T S_W W\right| = 0$ and $\left|W^T S_B W\right| \neq 0$.

$$W_{DCV} = argmax^T_{|W^T S_W W|=0} \left|W^T S_B W\right| \tag{3}$$

NLDA shows good performance especially when the number of input features of the data samples and the null space of S_W are large.

BDA is a modified form of LDA. Unlike LDA, which maximizes the distance of the mean values of classes in a multi-class classification problem, the BDA aims to classify one class of interest and the rest [31]. The BDA constructs a positive sample in the form of a normal distribution, and negative samples constitute a feature space that is distributed away from the mean of positive samples, and has the following objective function. Assuming that (i) the data samples $\mathbf{x}^P$ and $\mathbf{x}^N$ are positive and negative samples, respectively, (ii) their numbers are N_P

and N_N, respectively, and (iii) the average of the positive samples is $\boldsymbol{\mu}^P$, the scatter matrix of the positive samples S_P and the scattering matrix for the negative samples are defined as shown in Eq. (4). The objective function of BDA is defined as shown in Eq. (5).

$$\begin{aligned} S_P &= \sum_{k=1}^{N_P} (\mathbf{x}_k^P - \boldsymbol{\mu}^P)(\mathbf{x}_k^P - \boldsymbol{\mu}^P)^T \\ S_N &= \sum_{k=1}^{N_N} (\mathbf{x}_k^N - \boldsymbol{\mu}^P)(\mathbf{x}_k^N - \boldsymbol{\mu}^P)^T \end{aligned} \tag{4}$$

$$W_{BDA} = argmax_W \frac{\left|W^T S_N W\right|}{|W^T S_P W|} \tag{5}$$

To avoid the small sample size problem [11], we use ν and γ instead of $S_N^R = (1-\nu)S_N + \frac{\nu}{n} tr[S_N I]$ and $S_P^R = (1-\gamma)S_P + \frac{\gamma}{n} tr[S_P I]$ by using a regularization factor S_N and S_P for each scattering matrix [31]. After investigating classification rates for various values of ν and γ, we set ν and γ to 0 and 0.1, respectively.

In the subspace-based analyses, after finding W in the training phase, the feature vector ($\mathbf{y} \in R^{m\times 1}$, $m < n$) for a given sample $\mathbf{x}$ can be obtained through a linear transformation as $\mathbf{y} = W^T\mathbf{x}$. Also, the problem is effectively solved by defining the covariance matrices and objective function according to the particular type of problem. However, the number of input features should be secured for the covariance analysis of the input features to be successful. Besides, some methods, such as NLDA, may not be able to conduct an analysis if the number of input features is less than the number of samples. Therefore, to more efficiently use subspace discriminant analysis, it is necessary to ensure a certain number of input features.

2.2 Classification performance over data dimensionality

To confirm the effect of the dimension of the data sample on the classification performance in the subspace discriminant analysis, it is necessary to examine how the classification rate changes with respect to the data representing the same object with vectors of different dimensions. As an example, we performed recognition experiments on facial images with various resolutions [6]. We have experimented on images with 120×100, 60×50, 30×25, 24×20 and 15×12 resolution for the FERET database [25], CMU-PIE database [1], Yale B [12] and Yale database [32] database (Fig. 1). The NLDA method was used for 120×100, 60×50, 30×25, and 24×20 images, and the LDA method [2] was used for 15×12 images because there is no null space of S_W.

As can be seen in Fig. 2, the recognition rate decreases as the resolution decreases in all databases. The recognition rate of the 15×12 images, which can not use the NLDA method, is significantly lower than that for the 120×100 to 24×20 images because the applicable classification methods are limited when the dimension of the data is low. As a result, when data is a dimension higher than a certain level, it is possible to attempt effective classification using various methods. On the other hand, the variations in illumination and facial expression in facial images can be regarded as a kind of noise. In this sense, the FERET database, which has less variation in images than the CMU-PIE, Yale B and Yale databases, can be regarded as relatively noiseless.

The results of the experiment for the FERET database show that the recognition rates for 120×100, 60×50 and 24×20 images are almost the same. This indicates that, when the influence of the noise is not large, if the dimension of the data becomes larger than a certain level, there is no further advantage in classification accuracy, and the amount of unnecessary calculation

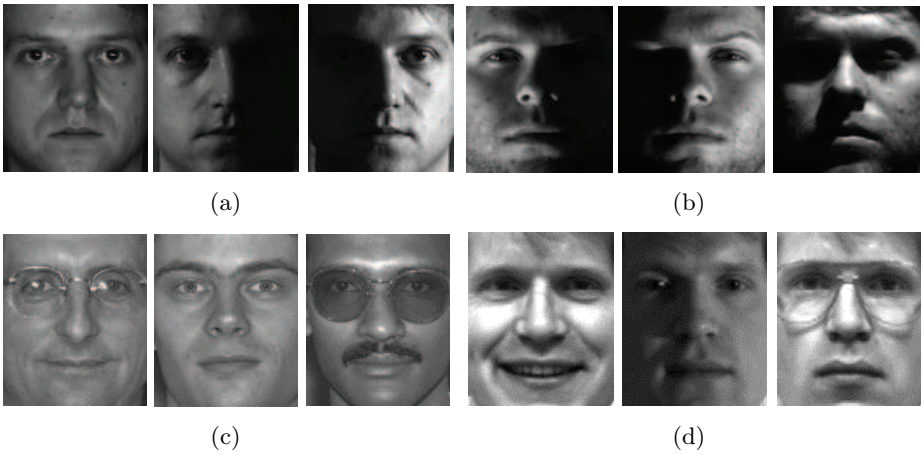

Figure 1: Examples from (a) CMU-PIE database. (b) Yale B database. (c) FERET database. (d) Yale database.

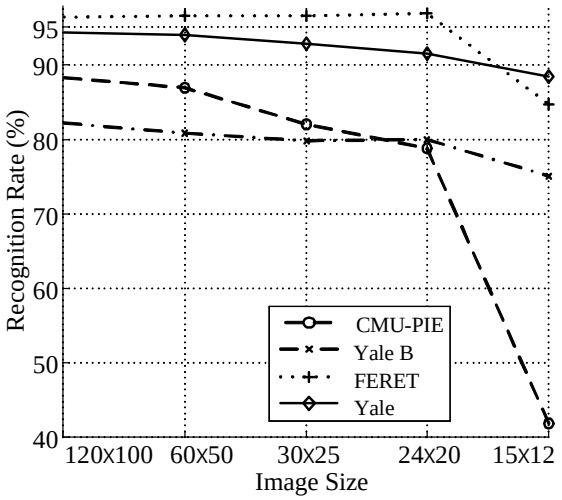

Figure 2: Face recognition rates for various face image resolutions.

increases due to data redundancy. Therefore, to efficiently classify the data, it is necessary to construct appropriately sized data.

3 Feature generation and construction of optimal features

As noted above, low-dimensional data samples may have limitations when classified only with the original input features. Therefore, to improve the performance of the data classification, it is desirable to increase the separability of the samples by converting the dataset with the samples into a high-dimensional space through a non-linear transformation $\varphi(\cdot)$ (Cover' theorem [10], Fig. 3). One simple way to increase the dimension of the input feature space is to create and add a higher order term from the input features of the data sample.
In this paper, we use the correlation between the input features as a new feature by adding the quadratic term $(x_i x_j, (i,j = 1,...,n))$ of the input features (pixels) of the data sample $(\mathbf{x} = [x_1,...,x_n]^T)$. The dimension of the data increases through the addition of a higher order n_{new} as follows.

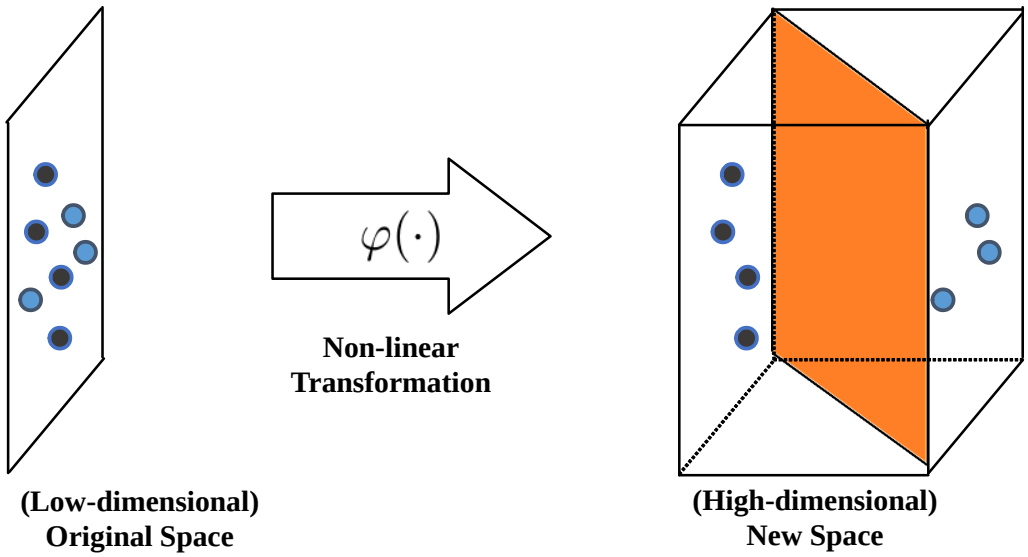

Figure 3: Cover's theorem.

$$n_{new} = \binom{n+2-1}{2} + n = \frac{(n+2-1)\,!}{2\,!(n-1)\,!} + n = \frac{(n^2+3n)}{2} \tag{6}$$

Since the dimension of the input feature space increases through feature generation, the accuracy of the whole classification can be improved. However, at the same time, the amount of computation needed in the classifier increases due to the increase in dimensionality. Furthermore, if the dimension of the feature space increases beyond a certain level, the classification accuracy would be rather reduced due to overfitting or the like. This phenomenon is called the "curse of dimensionality" [23]. This happens because as the dimension of the feature vector increases, the volume of the feature space increases exponentially, so the number of data samples required to effectively utilize the huge feature space also increases. However, there is a limit to collecting the necessary data samples in reality.

Since all generated input features do not have a positive effect on the classification performance, creating a feature is not itself a solution to the problem. For example, for an image with a size of 100×120, according to Eq. (6), 16290 input features can be created by adding a quadratic term, and some of these features are useful for classification, while others have little effect in solving the classification problems. Therefore, to obtain the optimal classification performance, it is necessary to generate only useful input features to construct a new input feature space of the appropriate dimensions.

Using the proposed method, the amount of discriminative information of individual features is quantitatively measured before using the original input features and the generated input features in the classification process. Then, based on the results of the measurement, only features with a large amount of discriminative information are selected to construct a new input feature vector, and the discriminant features that are to be used for classification are extracted using subspace discriminant analysis on the new input feature vectors $\mathbf{x}^{SFG}$.

The separability of the individual features is measured using the discriminant distance scale [21]. The distance between the different classes and the class can be defined as follows for a j-th component (feature) of $\mathbf{x}^{FG}$, where $\mathbf{x}^{FG} \in R^{n_{new}\times 1}$ is a data sample including newly generated input features.

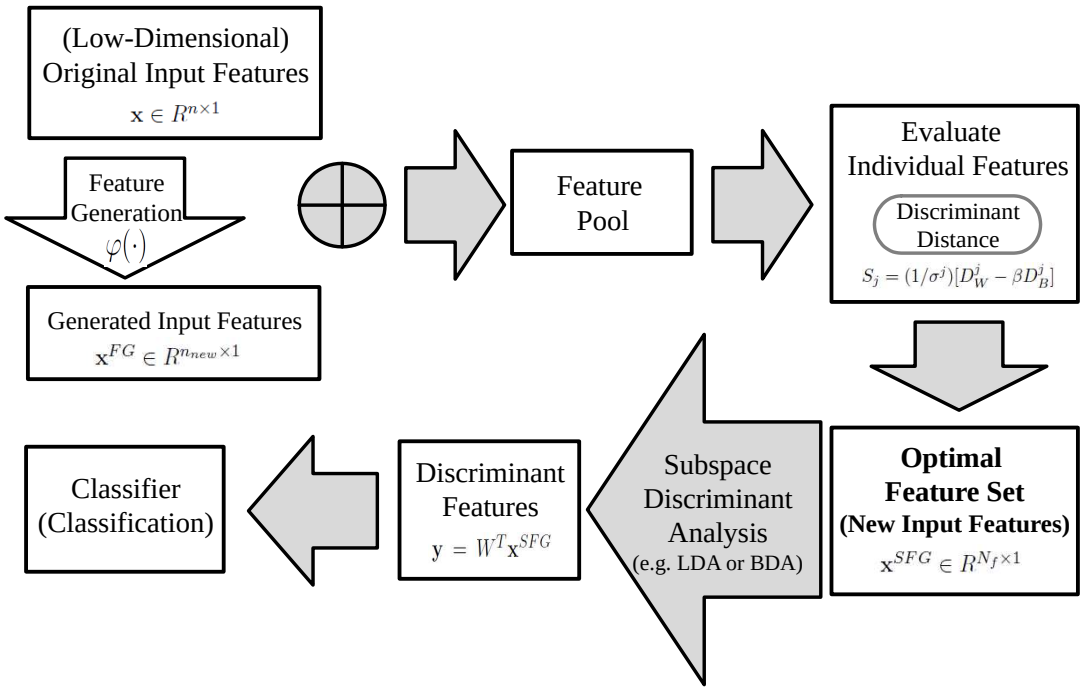

Figure 4: Overall procedure of the proposed method.

$$
\begin{aligned}
D_W^j &= \sum_{i=1}^{C} \frac{N_i}{N(N_i-1)} \sum_{x_{kj}^{FG}\in c_i} \left(x_{kj}^{FG} - \mu_i^j\right)^2 \\
D_B^j &= \sum_{i=1}^{C} \frac{N_i}{N} (\mu_i^j - \mu^j)^2
\end{aligned}
\tag{7}
$$

, where $\boldsymbol{\mu}_i^j$ and $\boldsymbol{\mu}^j$ are the j-th component of the mean of class c_i and all training data samples, respectively. The discriminant distance of the j-th feature from Eq. (7) can be defined as $(1/\sigma^j)[D_B^j - \beta D_W^j]$, $\sigma^j = (1/(N-1))\sum_{k=1}^{N}(x_{kj}^{FG} - m^j)^2$ [21], which can be used as a measure of the amount of discriminative information possessed by the j-th feature. β can be determined according to the distribution of the data samples as a user coefficient. In the case where the distribution within a class is large but the class separability is relatively good, it is preferable to reduce the value of β, which means a penalty of D_W^j. We set the value of β to 2 in this paper. Then, a measurement vector $\mathbf{S} = [S_1, S_1, .., S_{n_{new}}]^T$, $S_j = (1/\sigma^j)[D_W^j - \beta D_B^j]$ of the same size as n_{new} is defined and the new input feature vector $\mathbf{x}^{SFG}$ is constructed with the features corresponding to the large S_j. The entire process of the proposed method is shown in Fig. 4.

4 Experimental results and discussion

To show the effectiveness of the proposed method, we applied the proposed method to various real world problems. Through experiments on face databases and the UCI machine learning repository [3], we show that the proposed method works effectively for various kinds of low-dimensional data sets and for low-resolution images.

Table 1: Datasets from UCI machine learning repository used in the experiments

Dataset	**No. of classes**	**No. of instances**	**No. of original input f.**	**No. of new input f.** (for LDA/BDA)
Breast cancer	2	683	9	40/6
Pima	2	768	8	27/9
Bupa	2	345	6	20/15
Monk3	2	432	6	4/4
Balance	3	625	4	6/2
Wine	3	178	13	28/61
Glass	6	214	9	6/21
Car	4	1728	6	13/20

Table 2: Classification rates for UCI data sets

Feature extraction	**LDA**					**BDA**				
DatasetInput features	$\mathbf{x}^{ori}$	$\mathbf{x}^{IVS}$	$\mathbf{x}^{FG}$	$\mathbf{x}^{com}$	$\mathbf{x}^{SFG}$	$\mathbf{x}^{ori}$	$\mathbf{x}^{IVS}$	$\mathbf{x}^{FG}$	$\mathbf{x}^{com}$	$\mathbf{x}^{SFG}$
Breast.	95.9	96.0	95.7	96.5	96.0	95.1	95.8	95.3	96.8	95.8
Pima	68.9	69.1	69.8	68.6	**70.7**	69.3	70.0	69.5	68.7	**70.4**
Bupa	59.8	59.8	63.7	57.7	**64.1**	64.1	65.5	62.7	63.7	**64.8**
Monk3	87.4	100.0	91.2	99.6	**99.9**	68.6	100	68.4	99.4	**99.8**
Balance	87.7	87.7	94.1	88.9	**99.2**	84.3	84.3	85.3	96.3	**99.8**
Wine	98.7	98.7	96.4	98.7	98.6	98.0	98.8	99.3	98.6	**99.7**
Glass	61.8	71.2	64.5	71.7	71.5	71.2	77.6	70.0	72.3	70.6
Car	83.5	90.7	91.9	87.3	**94.9**	95.3	95.3	95.3	87.0	**95.5**
aver.	80.4	84.1	83.4	83.6	**86.8**	80.7	85.9	80.7	85.3	**87.0**

4.1 UCI Machine learning repository

We applied the propose method to several data sets in UCI machine learning repositories. Brief summaries of eight data sets that have been used in many other studies are given in Table 1. For each data set, we performed 10-fold cross validation 10 times and computed the average classification rate. Each input feature in the training set was normalized to have zero mean and unit variance, and the input features in the test set were also normalized using the means and variances of the training set. The one nearest neighbor rule was used as a classifier and the l_2 norm was used to measure the distance between two samples.

LDA and BDA were used to extract the discriminant features from the input feature vectors. LDA is a supervised learning method that is extensively used in data classification. In addition, as shown in Table 1, most of the data sets used in the experiments have binary classes, so we evaluated the classification performance using the BDA developed for one-class problems as well. We should find ways to extend BDA to multi-class problems in order to apply it to a few data sets having more than two classes, such as an iris data set, balance data set, glass data set and car data set. One of the simplest ways [20] to extend the BDA to D-class classification problems is to construct D data sets with only two classes (positive and negative). In constructing the i-th data set, the samples from the i-th class are regarded as positive samples, and the rest are regarded as negative samples. Then, we obtain D feature spaces by applying BDA to each of these data sets. During the test of a sample, a combined feature vector, which is concatenated with D resulting feature vectors from D feature spaces as in [19] is used with the classifier. The necessary parameters for CLDA and CBDA, i.e., the length of a composite vector and the number of composite features, were set to the values with which each classification method exhibited the best performance, as in [17].

Table 2 shows the classification performance using LDA and BDA for new input feature vectors obtained by applying various methods to input features. The values in the column corresponding to $\mathbf{x}^{ori}$ are the classification rates obtained by applying LDA or BDA to the original data. $\mathbf{x}^{IVS}$ are the data samples containing only some input features selected by the IVS method [8] among the original input features, and $\mathbf{x}^{FG}$ are data samples with quadratic terms added to original input features using Eq. (6). Columns corresponding to $\mathbf{x}^{Com}$ are the results of CLDA and CBDA using a composite vector, which is a subset of input features. For the last row, the average classification rate of nine data sets was reported for each method.

From the results in the table, the proposed method that selectively generated new input features ($\mathbf{x}^{SFG}$) provided the best classification performance in most data sets, showing that the average classification rates were 6.% and 6.3% higher in the LDA and BDA, respectively, than when using the original input features. The effects of the proposed method are prominent in the monk3 and balance data sets. In particular, for the balance data set, both the LDA and BDA classification results showed that when new features were selectively generated using the proposed method, the classification rate increased by more than 10% when using the original input features intact. The common characteristic of these two data sets is that the input features have fewer types of values. In both the monk3 and balance data sets, input feature values can only be four and five kinds of integers, respectively. In this case, when new features are generated using the proposed method, not only the dimension of the data but also the kinds of values that the input feature can have increases, so the data samples can be distributed more effectively in the feature space. On the other hand, in the case of the monk3 data set, LDA and BDA showed 87.4% and 68.6% of the original input features, respectively. However, when some input features were removed using the IVS method, both LDA and BDA showed 100%, respectively. This means that among the original input features, unnecessary features were included that would disturb the classification. As a result, the performance of $\mathbf{x}^{FG}$, including all quadratic terms generated by these unnecessary input features, increased slightly (in the case of LDA) or was even lower than the for $\mathbf{x}^{ori}$ (in the case of BDA). However, in the case of the selective feature generation using the proposed method ($\mathbf{x}^{SFG}$), the classification rate can be seen to have increased to nearly 100% because the unnecessary input features were effectively filtered.

4.2 Face database and preprocessing

We also applied the proposed method to a face recognition problem. The FERET, CMU-PIE database, Yale B, and Yale databases, which are used in the experiments, are widely used in face recognition research (Table 3, Fig. 2). In order to represent each database's degree of variation, we selected an image taken under normal conditions (no illumination and expression variations) for each subject as a reference image and computed the PSNR of the subject's other images. As shown in Table 3, the PSNR of the FERET database is higher compared to the other databases; thus, the images in the FERET database exhibited a relatively small variation.

For the FERET database, images for 792 subjects were used, and two images ('fa', 'fb') taken from the front of each person were used, that is, a total of 1584 images [7]. Of 792 participants, 200 images for 100 subjects were used as training images to evaluate the recognition performance, and the remaining images for 692 subjects were used as test images. For the test, the 'fa' image was used as a gallery image and the 'fb' image was used as the probe image.

Among the frontal pose images of the CMU-PIE database, the 'illum' category includes 21 images with different lighting conditions for a total of 68 subjects. In this experiment, we used 21 images for 65 subjects, that is, 1365 images in total, except for images of people who have some shooting defects or do not include all 21 kinds of illumination variations. We used three images ('27_06', '27_07', '27_08') for each subject, i.e., 195 total images that have a relatively

Table 3: Characteristics of each face database used for the experiments

Database	FERET	CMU-PIE	Yale B	Yale
No. of subjects	992	65	10	15
No. of images per subject	2	21	45	11
Illumination variation	none	large	large	small
Expression variation	small	none	none	large
Occlusion	none	none	none	glasses
No. training / test	200 / 1384	195 / 1170	70 / 380	10-fold CV
Degree of variations (avr.PSNR)	16.9	12.6	12.4	14.1

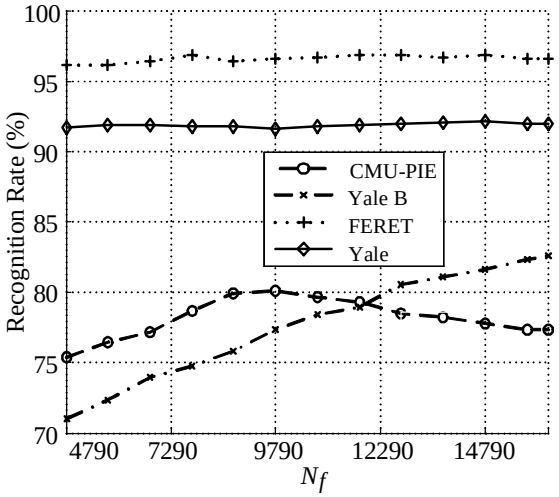

Figure 5: Recognition performance for various N_f.

small variation in illumination as training images, and the '27_20' image from the front lighting was used as a gallery image. The remaining images for each subject (total 65 pieces x17 = 1105 pieces) were used as proof images.

The Yale B database contains images for 10 subjects, and each subject's image consists of 45 kinds of images with illumination variations. The images are divided into subsets 1, 2, 3, and 4 according to the degree of variation in the illumination. In this experiment, the images for the subset 1 with less variation in illumination were used as training images and gallery images, and the images for remaining subset 2, 3 and 4 were used as probe images.

The Yale database contains 165 gray images of 15 subjects, with different facial expressions, with or without glasses, and under different illumination variations. In order to evaluate the recognition rates, we performed 10-fold cross validation 10 times and computed the average classification rate.

For face recognition experiments, facial images should be aligned to have the same size. For this, the whole face image is cropped based on the distance between the two eyes using manually detected eye coordinates and is then down scaled to a size of 120×100 [8], and the 60×50, 30×25, 24×20, and 15×12 images are downscaled versions of the 120×100 image again. All images were pre-processed for histogram equalization [13] and all pixels were normalized to have zero mean and unit standard deviation [7, 8]. The face recognition rates were evaluated from the 15×12 image (I_{180}), for which the recognition rate decreased sharply in Fig. 2, to I^{IP}_{12000} which is resized from the I_{180} to the 120x100 size via the bicubic interpolation [16], I^{FG}, to which the features

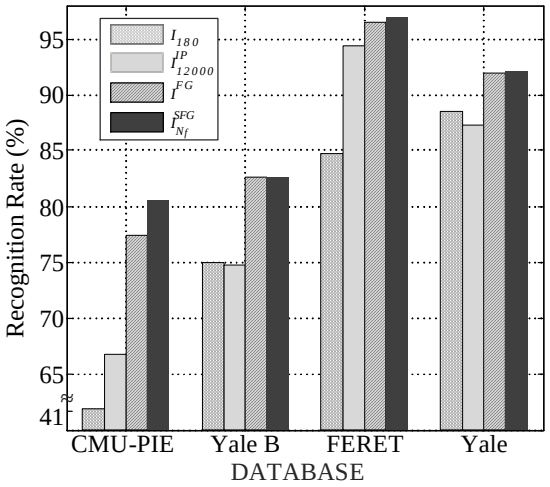

Figure 6: Comparison of recognition rates for I_{180}, I^{IP}_{12000}, I^{FG} and $I^{SFG}_{N_f}$(proposed method).

generated using Eq. (6) from I_{180} are added. $I^{SFG}_{N_f}$ (proposed) consists of the optimal features selected by the discriminant distance scale from 16,290 features of I^{FG}. The optimal number of features (N_f) is experimentally determined because it depends on the nature of the database [8]. As in Fig. 5, after we investigated the recognition rates by changing N_f, we set N_f to 12290, 10290, 16290, and 14290 for the FERET, CMU-PIE, Yale B, and Yale databases, respectively. Among the appearance-based face recognition methods, the DCV method was used for feature extraction and the Fisherface method was used only for I_{180} where the SSS problem occurred. The NN (Nearest Neighborhood) method was used as a classifier, and the Euclidean distance was used as the distance measurement.

Fig. 6 shows the recognition rates for I_{180}, I^{IP}_{12000}, I^{FG}, $I^{SFG}_{N_f}$ for various databases. Fig. 6 shows that the recognition rate for the CMU-PIE database and the FERET database was improved when the image size (i.e., the number of pixels) was increased using the interpolation method (I^{IP}_{12000}), but in the case of the Yale database, the recognition rate for I^{IP}_{12000} is less than I_{180} because the pixels (input features) generated via the interpolation method have brightness values estimated from the spatial relationship of adjacent pixels in the existing image, and thus the generated pixels do not help extract features using linear discriminant analysis. On the other hand, $I^{SFG}_{N_f}$, which is composed of the selected features by the discriminant distance scale among features generated in a non-linear way, showed a higher recognition rate than I_{180} for all databases.

Compared to I^{IP}_{12000}, the recognition rates of $I^{SFG}_{N_f}$ were significantly improved in the CMU-PIE, Yale B, and Yale databases than in the FERET database. The images of the FERET database, which have a relatively small variation compared to the CMU-PIE, Yale B and Yale databases, are less likely to suffer a loss of identity information due to image reduction. Since the images of the CMU-PIE, Yale B and Yale databases have already lost much of the identity information in the original image due to the variations such as in illumination and facial expressions, the reduced image (I_{180}) includes many pieces of face identification information as well as distortion information. Bicubic interpolation uses 16 adjacent pixels in I_{180} to determine the brightness value of a new pixel when expanded from I_{180} to I^{IP}_{12000}, so if any one of the 16 pixels contains distorted information (variation), the distortion is also reflected in the generated pixels. Consequently, in the case of the CMU-PIE, Yale B, and Yale databases, the improvement in the recognition rate through the use of I^{IP}_{12000} is not large or is rather worse than using I_{180}. In

contrast, the features generated by using the high order terms of the input features are relatively low in the distortion ratio of the identity information, and as a result, the recognition rate of I^{FG} is higher than I_{180} in all databases. In addition, even if distorted information is included in the generated features, all features are evaluated using the discriminant distance scale. Using only features with a high separability based on this ($I_{N_f}^{SFG}$), an additional improvement in the recognition rate can be obtained.

5 Conclusions

In pattern recognition problems, data for an object is represented vector composed of input features. The dimensions of data sample are determined by the attributes of the object samples are often stored as low-dimensional vectors according to the nature of the problem. Several discriminant feature extraction methods developed for data classification use statistical correlation of input features, but their performance is limited when the dimension of data is small or the range of values input features is small. Also, in the case of high-dimensional data such as image data, the image taken from a high-resolution camera converted into low-resolution image to reduce the calculation for data processing and effectively use the storage space. However, the performance may when a low-resolution image is used for recognition due to loss of information occur reducing the dimension of data. In this paper, we propose an input feature generation method effectively low-dimensional data to solve these problems. First, by generating high-order terms of the input features of the low-dimensional data samples, information on the correlation between the input features used as a new feature candidate group. Then, using the discriminant distance scale, new data samples were constructed with only input features with high separability by removing the features that are not helpful or obstructive to classification among the original input features and newly generated features. The experimental results on various low-dimensional data sets of UCI machine learning repository and several kinds of low-resolution facial images showed that the classification performance improved by selectively generating input features using the proposed method.

Acknowledgments

This work was supported by the Human Resources Program in Energy Technology of the Korea Institute of Energy Technology Evaluation and Planning (KETEP) granted financial resource from the Ministry of Trade, Industry and Energy, Republic of Korea (no. 20174030201740), and also supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2017-2015-0-00363) supervised by the IITP (Institute for Information and communications Technology Promotion).

Author's contributions

Conceived and designed the experiments: S.-I. Choi (SIC), H. Yoo (HY), S.T. Choi (STC). Performed the experiments: SIC, STC. Analyzed the data: SIC, HY, STC. Contributed reagents/-materials/analysis tools: SIC, HY. Wrote the paper: SIC, HY. Revised the manuscript critically for important intellectual content: SIC, HY, STC.

Bibliography

[1] Baker, S.; Sim, T.; Bsat, M. (2003); The CMU pose, illumination, and expression database, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, DOI: 10.1109/TPAMI.2003.1251154, 25(12), 1615-1618, 2003.

[2] Belhumeur, P. N.; Hespanha, J. P.; Kriegman, D. J. (1997); Eigenfaces vs. fisherfaces: Recognition using class specific linear projection, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, DOI: 10.1109/34.598228, 19(7), 711-720, 1997.

[3] Blake, C.; Merz, C. J. (1998); UCI Repository of machine learning databases, https://www.nist.gov/, 1998.

[4] Cevikalp, H.; Neamtu, M.; Wilkes, M.; Barkana, A. (2005); Discriminative common vectors for face recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, DOI: 10.1109/TPAMI.2005.9, 27(1), 4-13, 2005.

[5] Chen, S.; Zhu, Y.; Zhang, D.; Yang, J.-Y. (2005); Feature extraction approaches based on matrix pattern: MatPCA and MatFLDA, *Pattern Recognition Letters*, DOI: 10.1016/j.patrec.2004.10.009, 26(8), 1157-1167, 2005.

[6] Choi, S.-I. (2015); Feature generation method for low-resolution face recognition, *Journal of Korea Multimedia Society*, 18(9):1039-1046, 2015.

[7] Choi, S.-I.; Choi, C.-H.; Jeong, G.-M.; Kwak, N. (2012); Pixel selection based on discriminant features with application to face recognition, *Pattern Recognition Letters*, DOI: 10.1016/j.patrec.2012.01.005, 33(9), 1083-1092, 2012.

[8] Choi, S.-I.; Oh, J.; Choi, C.-H.; Kim, C. (2012); Input variable selection for feature extraction in classification problems, *Signal Processing*, ISSN: 01651684, DOI: 10.1016/j.sigpro.2011.08.023, 92(3), 636-648, 2012.

[9] Cortes, C.; Vapnik, V. (1995); Support-vector networks, *Machine Learning*, DOI: 10.1023/A:1022627411411, 20(3), 273-297, 1995.

[10] Cover, T. M. (1965); Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition, *IEEE Transactions on Electronic Computers*, ISSN: 03677508, DOI: 10.1109/PGEC.1965.264137, (3):326-334, 1965.

[11] Duda, R. O.; Hart, P. E.; Stork, D. G. (2001); *Pattern classification. 2nd*, New York, 55, 2001.

[12] Georghiades, A. S.; Belhumeur, P. N.; Kriegman, D. J. (2001); From few to many: Illumination cone models for face recognition under variable lighting and pose, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, DOI: 10.1109/34.927464, 23(6), 643-660, 2001.

[13] Gonzalez, R.; Woods, R. (2002); *Digital image processing*, A. Dowrkin, Ed. Upper Saddle River, New Jersey 07458, Prentice Hall, 2002.

[14] Jain, A. K.; Duin, R. P. W.; Mao, J. (2000); Statistical pattern recognition: A review, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, I, DOI: 10.1109/34.824819, 22(1), 4-37, 2000.

[15] Keinosuke, F. (1990); *Introduction to statistical pattern recognition*, Academic Press Inc., 1990.

[16] Keys, R. (1981); Cubic convolution interpolation for digital image processing, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, I, DOI: 10.1109/TASSP.1981.1163711, 29(6), 1153-1160, 1981.

[17] Kim, C. (2007); *Pattern recognition using composite features*, Ph. D. Thesis, Seoul National University, 2007.

[18] Kim, C.; Choi, C.-H. (2007); A discriminant analysis using composite features for classification problems, *Pattern Recognition*, DOI: 10.1016/j.patcog.2007.02.008, 40(11), 2958-2966, 2007.

[19] Kim, C.; Oh, J. Y.; Choi, C.-H. (2005); Combined subspace method using global and local features for face recognition, *Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on*, DOI: 10.1109/IJCNN.2005.1556212, 4, 2030-2035, 2005.

[20] Kwak, N.; Oh, J. (2009); Feature extraction for one-class classification problems: Enhancements to biased discriminant analysis, *Pattern Recognition*, I DOI: 10.1016/j.patcog.2008.07.002, 42(1), 17-26, 2009.

[21] Liang, J.; Yang, S.; Winstanley, A. (2008); Invariant optimal feature selection: A distance discriminant and feature ranking based solution, *Pattern Recognition*, DOI: 10.1016/j.patcog.2007.10.018, 41(5), 1429-1439, 2008.

[22] Lin, F.; Zhou, X.; Zeng, W. (2016); Sparse online learning for collaborative filtering, *International Journal of Computers Communications & Control*, 11(2), 248-258, 2016.

[23] Marimont, R.; Shapiro, M. (1979); Nearest neighbour searches and the curse of dimensionality, *IMA Journal of Applied Mathematics*, DOI: 10.1093/imamat/24.1.59, 24(1), 59-70, 1979.

[24] Mika, S.; Ratsch, G.; Weston, J.; Scholkopf, B.; Mullers, K.-R. (1999); Fisher discriminant analysis with kernels, *Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop*, 41-48, 1999.

[25] Phillips, P. J.; Wechsler, H.; Huang, J.; Rauss, P. J. (1998); The FERET database and evaluation procedure for face-recognition algorithms, *Image and vision computing*, 16(5), 295-306, 1998.

[26] Suto, J.; Oniga, S.; Pop Sitar, P. (2016); Feature analysis to human activity recognition, *International Journal of Computers Communications & Control*, ISSN: 18419836, 12(1), 116-130, 20106.

[27] Turk, M.; Pentland, A. (1991); Eigenfaces for recognition, *Journal of Cognitive Neuroscience*, DOI: 10.1162/jocn.1991.3.1.71, 3(1), 71-86, 1991.

[28] Viriri, S.; Lagerwall, B. (2016); Increasing face recognition rates using novel classification algorithms, *International Journal of Computers Communications & Control*, 11(3), 381-393, 2016.

[29] Xiong, H.; Swamy, M.; Ahmad, M.O. (2005); Two-dimensional FLD for face recognition, *Pattern Recognition*, ISSN: 00313203, DOI: 10.1016/j.patcog.2004.12.003, 38(7), 1121-1124, 2005.

[30] Yang, J.; Frangi, A. F.; Yang, J.-y.; Zhang, D.; Jin, Z. (2005); KPCA plus LDA: a complete kernel fisher discriminant framework for feature extraction and recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, DOI: 10.1109/TPAMI.2005.33, 27(2), 230-244, 2005.

[31] Zhou, X. S.; Huang, T. S. (2001); Small sample learning during multimedia retrieval using biasmap, *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, DOI: 10.1109/CVPR.2001.990450, 1, 111-117, 2001.

[32] Center for Computational Vision and Control, Yale University, The Yale FaceDatabase, http://cvc.cs.yale.edu/cvc/projects/yalefaces/yalefaces.html.