Journal of AI and Data Mining
Vol 7, No 1, 2019, 203-212 DOI: 10.22044/JADM.2017.5063.1610

2D Dimensionality Reduction Methods without Loss

S. Ahmadkhani[1], P. Adibi[2*] and A. Ahmadkhani[3]

1. Young Researchers & Elite Club, Kermanshah Branch, Islamic Azad University, Kermanshah, Iran
2. Department of Artificial Intelligence, Computer Engineering Faculty, University of Isfahan, Isfahan, Iran.
3. Department of Mechanical Engineering, Engineering Faculty,Razi University of Kermanshah, Kermanshah, Iran.

Received 19 November 2016; Received 27 February 2017; Accepted 11 November 2017
**Corresponding author: adibi@eng.ui.ac.ir (P. Adibi).*

Abstract
In this work, several 2D extensions of the principal component analysis (PCA) and linear discriminant analysis (LDA) techniques were applied in a lossless dimensionality reduction framework for face recognition applications. In this framework, the benefits of dimensionality reduction were used to improve the performance of its predictive model, which was a support vector machine (SVM) classifier. At the same time, the loss of useful information was minimized using the projection penalty idea. The well-known face databases were used to train and evaluate the proposed methods. The experimental results obtained indicated that the proposed methods had a higher average classification accuracy, in general, compared to the classification based on the Euclidean distance, and also compared to the methods that first extracted the features based on the dimensionality reduction technics, and then used the SVM classifier as the predictive model.

Keywords: *Lossless Dimensionality Reduction, Face Recognition, Support Vector Machine, 2DPCA, 2DLDA, $(2D)^2$PCA, $(2D)^2$LDA, Projection Penalty.*

1. Introduction

Dimensionality reduction is a key problem in the machine learning, pattern recognition, and computer vision fields. The learning of a classifier in high-dimensional spaces with a limited number of training examples is a difficult task. As it stands, many real-world problems have been defined in high-dimensional input spaces, and a lot of classification methods are limited and inefficient for high-dimensional data. As a result, many researchers have adopted various dimensionality reduction techniques to reduce the complexity of the problem by reducing the dimension of its feature space. This can reduce the computational cost of the next steps and enhance the overall system performance. Moreover, reducing the feature vector size using the dimensionality reduction techniques typically increases the classification accuracy with preventing the curse of dimensionality problem. A variety of ways for dimensionality reduction have been proposed and the practical importance of these methods have been studied extensively [1, 2].

The most famous dimensionality reduction method is probably the principal component analysis (PCA) technique. This is an unsupervised method that can find a number of eigenvectors of the empirical covariance matrix of data corresponding to the largest eigenvalues of this matrix. These vectors have been considered as the basis vectors of the principal sub-space of the data. Since PCA finds a global linear sub-space for data, its performance is limited to the data distributed on non-linear manifolds. Linear manifold topographic map (LMTM) [3] is another unsupervised method, trying to remove the above-mentioned limitations of data representation with learning several local linear manifolds. This method is a neural model for dimensionality reduction through a topology-preserving lattice, and a piecewise linear approximation of the data principal manifold have been obtained [3]. The nonlinear manifold learning methods such as

isomap [4], locally linear embedding (LLE) [5], Laplacian eigenmap [6], and maximum variance unfolding (MVU) [7] are also several other dimensionality reduction methods that learn the underlying nonlinear manifold of the data directly. The above-mentioned techniquess are all examples of unsupervised dimensionality reduction methods; however, the target values of the training samples can be used in the dimensionality reduction process for supervised problems such as data classification in order to improve the final performance of the system (e.g. the classification accuracy).

One of the most famous supervised dimensionality reduction methods is the linear discriminant analysis (LDA) technique [8]. The objective of LDA is to find a sub-space, where the projected samples from the same class are close to each other, while the projected samples from different classes are far from each other. As a result, LDA achieves maximum discrimination between classes in its lower-dimensional representation. LDA is a linear dimensionality reduction method, which works well only when the sample data is distributed on a linear sub-space in the original space. The Kernel discriminant analysis (KDA) [9] techniques have been proposed to dominate this linearly-distributed limitation. In addition, LDA algorithm has been developed with the assumption that each class of the samples has the same Gaussian distribution, a property that often does not exist in real-world applications. In the lack of this property, the separability of different classes cannot be well-characterized by the LDA algorithm. For this reason, the marginal Fisher analysis (MFA) method [10] has been proposed to overcome the problem corresponding to this data distribution assumption of the LDA using the graph embedding framework [10]. MFA cannot guarantee that the data neighbors after dimensionality reduction maintain the original structure. Therefore, the neighborhood preserving and marginal discriminant embedding (NP-MDE) [11] method has been proposed based on linear graph embedding (LGE) and MFA. NP-MDE minimizes the within-class scatter, while maximizes the margin among different classes. Moreover, the neighborhood structure with each class is preserved [11].

In [12] a supervised feature extraction method called DA-PC1 proposed, based on discriminant analysis (DA) which uses the first principal component (PC1). This method copes with the small sample size problem and has not the limitation of linear discriminant analysis (LDA) in the number of extracted features [12].

All the methods discussed above are based upon vectors analysis. Transforming the image matrices into image vectors should take place first when dealing with images. Then the optimal projection is obtained based on these vectors. There are methods that are directly based on the analysis of the original image matrices. For example, 2DPCA is based upon 2D matrices rather than 1D vectors [13], which means that the image matrix does not require to be converted into a vector. Therefore, 2DPCA has two advantages: 1) easier to evaluate the covariance matrix accurately and 2) lower time-consumption [14]. The feature fusion approach for 2DPCA [15] is another method based on the analysis of the original image matrices. This model achieves better recognition results with combining the features generated from the two schemes of 2DPCA. The SI2DPCA [16] approach, unlike the conventional 2DPCA, divides a whole image into smaller sub-images to increase the weights of the features resulting in a higher performance feature extraction. Meanwhile, the computational cost can also be reduced due to the smaller size of sub-images. 2DLDA is another method based on the analysis of the original image matrices instead of their 1D vector forms, which extends the original LDA idea [17]. The robust 2DPCA method has also been introduced for robust dimensionality reduction in the classification of 2D data in order to minimize the vector variance measure (MVV) [18]. The main drawback of the 2DPCA and 2DLDA methods is the necessity of huge feature matrices for the task of face recognition. In order to overcome this problem, the 2-directional 2DPCA [19] and 2-directional 2DLDA [20] methods have been developed.

The adjusted population value decomposition (APVD) technique has been proposed as a novel dimensionality reduction approach [21]. This model has a nice approximation property, and computes fast the reduction of high-dimensional data and explicitly incorporates the intrinsic 2D structure of the matrices.

Notwithstanding the advantages mentioned in the dimensionality reduction process, if this process is applied and then learning is performed in the reduced-dimensional space, the prediction performance can be degraded since any reduction loses information. Indeed, it is difficult to tell whether the information lost due to a dimensionality reduction procedure is relevant to a prediction task or not. Thus reducing the dimensionality of the feature space can be viewed

as the restriction of the model search to a parameter sub-space [21]. The basic idea in [21] is that instead of restricting the model search to a parameter sub-space, still search in the full parameter space can go on while the projection distance to this sub-space is penalized. As a result, dimensionality reduction has been used to guide the model search in the parameter space rather than restricting it. In [23], a supervised version for the probabilistic PCA mixture model (PPCAMM) called supervised PPCAMM (SPPCAMM) has been proposed. This algorithm that has been used in learning a predictive model with projection penalties [22] enables the model to gain from the dimensionality reduction techniques without losing relevant information.

In this work, the 2DLDA [17], $(2D)^2$LDA [20], 2DPCA[13], and $(2D)^2$PCA [19] methods were applied based on 2D image matrices, and then the feature matrices obtained were converted to feature vectors to be applied to the learning of a predictive model with projection penalties idea, enabling the model to gain from dimensionality reduction techniques without losing the relevant information. As it can be observed in the experimental results on the face datasets, the proposed methods have the best average classification accuracies compared to the other relevant dimensionality reduction methods.

This article has been organized as what follows. In Section 2, the 2DPCA, 2DLDA, $(2D)^2$PCA, and $(2D)^2$LDA dimensionality reduction methods and dimensionality reduction without loss of approach have been reviewed. In Section 3, the proposed methods, i.e. using 2D extensions of PCA and LDA for dimensionality reduction with minimum loss of information using SVM classifier as the predictive model have been explained. In Section 4, the experimental results on the well-known face recognition datasets have been reported. The conclusions have been presented in Section 5.

2. Related concepts

2.1. Two-dimensional PCA (2DPCA) and two-directional two-dimensional PCA ($(2D)^2$PCA)

Suppose $\{\mathbf{X}_i\}$ to be the set of training images, where the j'th sample is denoted by the $m\times n$ matrix $\mathbf{X}_j$. Let $\mathbf{a}$ to denote an n-dimensional unitary column vector. The idea of 2DPCA is to project image $\mathbf{X}$, an $m\times n$ random matrix, onto $\mathbf{a}$ by the linear transformation $\mathbf{y}=\mathbf{X}\mathbf{a}$ [13].

Thus we obtain an *m*-dimensional projected vector $\mathbf{y}$ called the projected feature vector of image $\mathbf{X}$.

Suppose that there are N training image samples in total; the j'th training image is denoted by an $m\times n$ matrix $\mathbf{X}_j$ $(j=1,2,...,N)$, and the average image of all training samples is denoted by $\overline{\mathbf{X}}$. The matrix $\mathbf{G}_t$, called the image covariance matrix, is calculated as follows [13]:

$$\mathbf{G}_t=\frac{1}{N}\sum_{j=1}^{N}(\mathbf{X}_j-\overline{\mathbf{X}})^T(\mathbf{X}_j-\overline{\mathbf{X}}) \tag{1}$$

The 2DPCA method selects a set of projection axes $\{\mathbf{a}_1,...,\mathbf{a}_d\}$ to maximize the criterion $J(\mathbf{a})=\mathbf{a}^T\mathbf{G}_t\mathbf{a}$. The optimal projection axes, subject to the orthonormal constraints, maximizes the criterion $J(\mathbf{a})$, as follows:

$$\begin{cases}\{\mathbf{a}_1,...,\mathbf{a}_d\}=\arg\max J(\mathbf{a})\\ \mathbf{a}_i^T\mathbf{a}_j=0,\quad i\neq j,\quad i,j=1,...,d\end{cases} \tag{2}$$

The optimal projection axes $\mathbf{B}_d=(\mathbf{a}_1,...,\mathbf{a}_d)$ are the orthonormal eigenvectors of $\mathbf{G}_t$ corresponding to the first *d* largest eigenvalues used for feature extraction. For a given image sample **X**, we have:

$$\mathbf{y}_k=\mathbf{X}\mathbf{a}_k,\ k=1,...,d \tag{3}$$

Then we obtain a set of projected feature vectors $\mathbf{y}_1,...,\mathbf{y}_d$ called the principal component vectors of the sample image **X**. It should be noted that each principal component of 2DPCA is a vector, whereas the principal components of PCA are scalars [13].

2DPCA learns the optimal projection matrix **B**, reflecting information between the rows of the image; then projects image **X** onto **B**, yielding an $m\times d$ feature matrix $\mathbf{Y}_{m\times d}=\mathbf{X}_{m\times n}.\mathbf{B}_{n\times d}$. If we consider the alternative 2DPCA [18], similarly, the optimal projection matrix **Z** reflecting information between columns of image is obtained; then **X** is projected onto **Z**, yielding a $q\times n$ feature matrix $\mathbf{T}_{q\times n}=\mathbf{Z}_{m\times q}^T\mathbf{X}_{m\times n}$.

$(2D)^2$PCA is a way to use the projection matrices **B** and **Z** simultaneously. This method projects the $m\times n$ image **X** onto **B** and **Z** simultaneously, yielding a $q\times d$ feature matrix **R** as follows [19]:

$$\mathbf{R}=\mathbf{Z}^T\mathbf{X}\mathbf{B} \tag{4}$$

2.2. Two-dimensional LDA (2DLDA) and two-directional two-dimensional LDA ((2D)2LDA)

Suppose $\{\mathbf{X}_{ij}\}$ is the set of training images in a classification problem that contains *c* classes. The *j*'th sample of the *i*'th class is denoted by the $m\times n$

matrix $\mathbf{X}_{ij}$, and the *i*'th class has N_i training samples. The total number of training samples is $N=\sum_{i=1}^{c}N_i$. The between-class scatter matrix $\mathbf{S}_b$ and within-class scatter matrix $\mathbf{S}_W$ are defined as follow [17, 24]:

$$\mathbf{S}_b=\sum_{i=1}^{c}N_i(\overline{\mathbf{X}}_i-\overline{\mathbf{X}})(\overline{\mathbf{X}}_i-\overline{\mathbf{X}})^T$$
$$\mathbf{S}_W=\sum_{i=1}^{c}\sum_{j=1}^{N_i}(\mathbf{X}_{ij}-\overline{\mathbf{X}}_i)(\mathbf{X}_{ij}-\overline{\mathbf{X}}_i)^T \quad (5)$$

In the above equations, $\overline{\mathbf{X}}$ and $\overline{\mathbf{X}}_i$ denote the mean of all samples and the mean of the *i*'th class samples, respectively.

The 2DLDA approach attempts to seek a set of projecting vectors $\{\mathbf{a}_1,...,\mathbf{a}_d\}$ that best discriminates different face classes with maximizing the criterion function $J(\mathbf{a})$ defined as follows:

$$J(\mathbf{a})=\frac{\mathbf{a}^T\mathbf{S}_b\mathbf{a}}{\mathbf{a}^T\mathbf{S}_W\mathbf{a}} \quad (6)$$

The vector $\mathbf{a}_{\text{opt}}$, which maximizes the above function, is called the optimal discriminant vector, and defined as follows:

$$\mathbf{a}_{opt}=\arg\max_{\mathbf{a}}J(\mathbf{a}) \quad (7)$$

If $\mathbf{S}_W$ is non-singular, the optimal vector of 2DLDA is the eigenvector corresponding to the maximal eigenvalue of the matrix $\mathbf{S}_W^{-1}\mathbf{S}_b$. Generally, the discriminant axes $\mathbf{B}_d=(\mathbf{a}_1,...,\mathbf{a}_d)$ are composed of orthogonal eigenvectors $\mathbf{a}_1,...,\mathbf{a}_d$ of $\mathbf{S}_W^{-1}\mathbf{S}_b$, corresponding to the first *d* largest eigenvalues. The feature matrix of $\mathbf{X}_{ij}$ is $\mathbf{Y}_{ij}=\mathbf{X}_{ij}\mathbf{B}$ obtained by projecting $\mathbf{X}_{ij}$ into the sub-space $\mathbf{B}$, and the size of $\mathbf{Y}_{ij}$ is $m\times d$ [24].

2DLDA works in the row-wise manner, considering the information between rows of the image to learn an optimal projection matrix **B**, and then project image **X** onto **B**, yielding an $m\times d$ feature matrix $\mathbf{Y}_{m\times d}=\mathbf{X}_{m\times n}.\mathbf{B}_{n\times d}$. Similarly, in the alternative 2DLDA [20], the optimal projection matrix **L** reflecting information between columns of the image is obtained, and then image **X** is projected onto **L**, yielding a $q\times n$ feature matrix $\mathbf{T}_{q\times n}=\mathbf{L}_{m\times q}^T\mathbf{X}_{m\times n}$.

$(2D)^2$LDA is a way to use the projection matrices **B** and **L** simultaneously. This method projects the $m\times n$ image **X** onto **B** and **L** simultaneously, yielding a $q\times d$ feature matrix **Q** as follows [20]:

$$\mathbf{Q}=\mathbf{L}^T\mathbf{X}\mathbf{B} \quad (8)$$

2.3. Dimensionality reduction without loss using projection penalty

Learning predictive models in high-dimensional spaces using dimensionality reduction techniques is popular; however, the predication performance can be degraded. In this section, we explain the projection penalties idea [22] that makes the predictive modeling to use dimensionality reduction without losing useful information. Consider the learning process of a linear prediction model with parameters **w** and *b* in a *p*-dimensional space by minimizing an empirical loss function *L* given a set of *N* training samples $\{(\mathbf{x}_i,t_i)\}_{i=1}^N$:

$$\underset{\mathbf{w}\in\mathrm{R}^p,b}{\operatorname{argmin}}\sum_{i=1}^{N}L(t_i,\mathbf{w}^T\mathbf{x}_i+b) \quad (9)$$

where, the weight vector $\mathbf{w}\in\mathrm{R}^p$ and the bias value *b* are used to represent the prediction model, and $\mathbf{x}_i$ and t_i are the *p*-dimensional feature vector and the response variable of the *i*'th example, respectively. The form of the empirical loss *L* depends on the choice of prediction model, e.g. a usual form is the squared error loss, and the other forms containing logistic log-likelihood or hinge loss. A linear dimensionality reduction in a *p*-dimensional input space can be represented by a $d \times p$ matrix **P**, where *d* is the dimension of reduced-dimensional space and $d < p$. For an input example **x** in the original feature space, **Px** is its representation in the reduced-dimensional space. In this sense, performing a linear dimensionality reduction and then learning a predictive model in reduced-dimensional space can be written as:

$$\underset{\mathbf{v}\in\mathrm{R}^d,b}{\operatorname{argmin}}\sum_{i=1}^{N}L(t_i,\mathbf{v}^T(\mathbf{P}\mathbf{x}_i)+b) \quad (10)$$

where, $\mathbf{v}\in\mathrm{R}^d$ is the parameter vector learned in the reduced-dimensional feature space.

Comparing (9) and (10) shows that performing a linear dimensionality reduction **P** in the feature space corresponds to confining the *p*-dimensional parameter vector **w** to a sub-space M_p defined as follows:

$$M_p=\left\{\mathbf{w}\in R^p \mid \mathbf{w}=\mathbf{P}^T\mathbf{v},\mathbf{v}\in R^d\right\} \quad (11)$$

Thus the model search is restricted to a parameter sub-space M_p. As a result, dimensionality reduction maybe lose information. Moreover, there is no guarantee that the optimal model

parameters in the reduced-dimensional space is the same as the optimal parameters in the original space. The basic idea of the projection penalty method is that search in the full parameter space can go on; however, the distance to the reduced-dimensional sub-space is penalized [22]. Using this idea, the predictive model can be written as [22]:

$$\underset{\tilde{\mathbf{w}}\in\mathrm{R}^{p},\mathbf{v}\in\mathrm{R}^{d},b}{\operatorname{argmin}} \sum_{i=1}^{N} L(t_i, \tilde{\mathbf{w}}^T\mathbf{x}_i + \mathbf{v}^T(\mathbf{P}\mathbf{x}_i) + b) + \lambda J(\tilde{\mathbf{w}}) \tag{12}$$

where, $\tilde{\mathbf{W}} = \mathbf{W} - \mathbf{P}^T\mathbf{v}$, $J(\cdot)$ is a penalty function such as $\|\cdot\|_2^2$ or $\|\cdot\|_1$, and λ is a regularization parameter. If we consider that both the predictive model and the dimensionality reduction operator have been designed to operate on a kernel feature space, and each training sample represented as $\Phi(\mathbf{x}_i)$ in the kernel feature space, then this prediction model with the projection penalty idea can be written as:

$$\underset{\mathbf{w}\in\mathrm{R}^{p},b}{\operatorname{argmin}} \sum_{i=1}^{N} L(t_i, \mathbf{w}^T\Phi(\mathbf{x}_i) + b) + \min_{\mathbf{v}\in\mathrm{R}^{d}} \lambda J(\mathbf{w} - \mathbf{P}^T\mathbf{v}) \tag{13}$$

Since the problem is classification, the hinge loss of SVM is used as the empirical loss $L(\cdot)$ and the $\ell 2$-norm penalty is used as $J(\cdot)$ in the above equation. Introducing slack variables $\{\xi_i\}_{i=1}^{N}$ for the hinge loss, we have the following quadratic programming problem:

$$\begin{aligned} &\underset{\mathbf{w}\in\mathrm{R}^{p},v\in R^{d},b,\{\xi_i\}_{i=1}^{N}}{\operatorname{argmin}} \frac{1}{2}\left\|\mathbf{w}-\mathbf{P}^T\mathbf{v}\right\|_2^2 + C\sum_{i=1}^{n}\xi_i \\ &\text{s.t. } t_i(\mathbf{w}^T\Phi(\mathbf{x}_i) + b) \geq 1-\xi_i \quad ,\forall i \\ &\qquad \xi_i \geq 0 \quad \forall i \end{aligned} \tag{14}$$

where, $C \propto \frac{1}{\lambda}$. Also using $\tilde{\mathbf{w}} = \mathbf{w} - \mathbf{P}^T\mathbf{v}$ [21]:

$$\begin{aligned} &\underset{\tilde{\mathbf{w}}\in\mathrm{R}^{p},v\in R^{d},b,\{\xi_i\}_{i=1}^{N}}{\operatorname{argmin}} \frac{1}{2}\left\|\tilde{\mathbf{w}}\right\|_2^2 + C\sum_{i=1}^{n}\xi_i \\ &\text{s.t. } t_i(\tilde{\mathbf{w}}^T\Phi(\mathbf{x}_i) + \mathbf{v}^T\mathbf{P}\Phi(\mathbf{x}_i) + b) \geq 1-\xi_i \quad ,\forall i \\ &\qquad \xi_i \geq 0 \quad \forall i \end{aligned} \tag{15}$$

In the above equation, both $\tilde{\mathbf{w}}^T\Phi(\mathbf{x}_i)$ and $\mathbf{P}\Phi(\mathbf{x}_i)$ have been computed via the kernel trick.

3. Proposed methods

In this article, we propose to use the 2D dimensionality reduction techniques in dimensionality reduction without a loss of framework. For this purpose, the 2DPCA, $(2D)^2$PCA, 2DLDA, and $(2D)^2$LDA methods were applied based on the 2D image matrices. The feature matrices obtained in these methods were converted into feature vectors $\mathbf{x}_i$ to be applied in the learning of common classifiers. Then these classifiers as predictive models with the projection penalty idea were used to reduce information loss. As seen in the previous section, the dimensionality reduction operator $\mathbf{P}$ is assumed to be either linear in the input space or at least linear in the kernel feature space. However, considering (12) and (15), the parts that involve the dimensionality reduction operator $\mathbf{P}$ are just $\mathbf{P}\mathbf{x}_i$ in (12) and $\mathbf{P}\Phi(\mathbf{x}_i)$ in (15). Therefore, a simple trick is to replace these terms with an arbitrary dimensionality reduction function $\Psi(\mathbf{x}_i)$ [22]. In this work, 2DPCA, $(2D)^2$PCA, 2DLDA, and $(2D)^2$LDA were used as the dimensionality reduction operator $\Psi(.)$. Replacing this function in (15) we have:

$$\begin{aligned} &\underset{\tilde{\mathbf{w}}\in\mathrm{R}^{p},\mathbf{v}\in R^{d},b,\{\xi_i\}_{i=1}^{N}}{\operatorname{argmin}} \frac{1}{2}\left\|\tilde{\mathbf{w}}\right\|_2^2 + C\sum_{i=1}^{n}\xi_i \\ &s.t. \quad y_i(\tilde{\mathbf{w}}^T\Phi(\mathbf{x}_i) + \mathbf{v}^T\Psi(\mathbf{x}_i) + b) \geq 1-\xi_i \quad ,\forall i \\ &\qquad \xi_i \geq 0 \quad \forall i \end{aligned} \tag{16}$$

In the proposed method, instead of reducing the dimensionality of the data by these two methods, and then searching for the parameters of predictive model in the learned reduced-dimensional spaces, which causes the loss of information, the search process was performed in the full parameter space; however, distances to the obtained lower-dimensional manifold were penalized.

4. Experimental results

In this section, the effectiveness of the proposed methods in face recognition application was tested. For this purpose, the Yale, ORL and Japanese Female Facial Expression (JAFFE) face datasets were used. The ORL database contains 10 different images from each of 40 distinct subjects. The images were the same size of 92×112 pixels, and for a number of subjects, the images were taken in different lighting conditions, facial expressions, and facial details. We considered all the 780 binary classification tasks, each classifying between two subjects. The average classification error over tasks was the performance measure, and the aggregated results over 10 runs were reported. The Yale face dataset contains 165 grayscale images of 15 individuals, with size 243×320 pixels. There are 11 images per subject, one per each facial configuration such as different expressions, emotions, illumination

conditions, and wearing glasses (or not). In this work, the normalized Yale face database was used, which contained rotated, cropped, and middle of eyes centered face images [25]. All the 105 binary classification tasks were considered, each classifying between two subjects. The average classification error over tasks is the performance measure, and the aggregated results over 10 runs were reported.

The JAFFE face dataset contains 213 images of 7 facial expressions (6 basic facial expressions + 1 neutral) posed by 10 Japanese female models, with size 256×256 pixels. From each subject, 21 images were used, and all 45 binary classification tasks, each classifying between two subjects, was considered. The average classification error over tasks is the performance measure, and the aggregated results over 10 runs were reported.

In order to evaluate the proposed methods, two types of tests using 2DPCA, $(2D)^2$PCA, 2DLDA, and $(2D)^2$LDA as feature extraction methods, and *k* nearest neighbor (*k*NN), SVM, and SVM with projection penalties as classifier were performed. In the experiments, three different sizes were considered for the training set, selecting 3, 5, and 7 images per subject. It is noteworthy that no additional pre-processing on the images of the datasets was made.

The first set of the experiments was performed and the average accuracy of the classification for each method on ORL, Yale, and JAFFE datasets were reported in tables 1, 2, and 3, respectively. The six experiments were set as follow: 1) first applying 2DPCA and then applying 1NN method with Euclidean distance metric for classification (the first column of Tables 1, 2, and 3); 2) first applying 2DPCA and then applying SVM for classification (the second column of Tables 1, 2, and 3); 3) first applying 2DPCA and then applying SVM with projection penalties for classification (the third column of Tables 1, 2, and 3); 4) first applying $(2D)^2$PCA and then applying 1NN method with Euclidean distance metric for classification (the forth column of Tables 1, 2, and 3); 5) first applying $(2D)^2$PCA and then applying SVM for classification (the fifth column of Tables 1, 2, and 3); 6) first applying $(2D)^2$PCA and then applying SVM with projection penalties for classification (the sixth column of Tables 1, 2, and 3). The proposed methods 2DPCA-ProjSVM and $(2D)^2$PCA-ProjSVM (the third and sixth columns of Tables 1, 2, and 3) in most of these experiments (especially for the ORL and JAFFE datasets) show the best accuracies.

Figures 1 and 2 show the average accuracy versus the number of dimensions in the reduced space, for the above mentioned six methods with 5 training images per subject, in 10 random runs on the ORL and JAFFE face datasets, respectively.

The resulting classification accuracies of the second set of the experiments on the ORL, Yale, and JAFFE datasets are reported in tables 4, 5, and 6, respectively. The six conducted experiments are set as follows: 1) first applying 2DLDA and then applying 1NN method with Euclidean distance metric for classification (the first column of Tables 4, 5, and 6). This is the same method introduced in [13]; 2) first applying 2DLDA and then applying SVM for classification (the second column of Tables 4, 5, and 6); 3) first applying 2DLDA and then applying SVM with projection penalties for classification (the third column of Tables 4, 5, and 6). 4) first applying $(2D)^2$LDA and then applying 1NN method with Euclidean distance metric for classification (the forth column of Tables 4, 5, and 6); 5) first applying $(2D)^2$LDA and then applying SVM for classification (the fifth column of Tables 4, 5, and 6); 6) first applying $(2D)^2$LDA and then applying SVM with projection penalties for classification (the sixth column of Tables 4, 5, and 6). The proposed methods 2DLDA-ProjSVM and $(2D)^2$LDA-ProjSVM (the third and sixth columns of Tables 1 and 2) in most of these experiments (for all datasets) show the best accuracies.

Table 1. Mean classification accuracy (in percent) over 10 random runs for ORL dataset with reduced dimension of 10 using 2DPCA and $(2D)^2$PCA as dimensionality reduction methods.

Training face (#)	2DPCA (112×10)	2DPCA-SVM (112×10)	2DPCA-ProjSVM (112×10)	$(2D)^2$PCA (10×10)	$(2D)^2$PCA-SVM (10×10)	$(2D)^2$PCA-ProjSVM (10×10)
3	88.64 ± 1.86	85.00 ± 1.81	**90.25 ± 1.67**	88.01 ± 1.95	83.14 ± 1.7	89.89 ± 1.60
5	94.45 ± 1.73	94.05 ± 1.86	96.10 ± 1.53	94.2 ± 1.74	92.45 ± 1.92	**96.10 ± 1.13**
7	96.75 ± 1.04	97.93 ± 1.37	97.75 ± 1.32	96.83 ± 1.07	96.58 ± 1.14	**97.93 ± 1.32**

Table 2. Mean classification accuracy (in percent) over 10 random runs for Yale dataset with reduced dimension of 10 using 2DPCA and (2D)^{2}PCA as dimensionality reduction methods.

Training face (#)	2DPCA (231×10)	2DPCA-SVM (231×10)	2DPCA-ProjSVM (231×10)	(2D)^{2}PCA (10×10)	(2D)^{2}PCA-SVM (10×10)	(2D)^{2}PCA-ProjSVM (10×10)
3	79.33 ± 2.25	81.67 ± 3.82	84 ± 2.90	80.17 ± 1.92	**90.08 ± 3.15**	84.58 ± 2.81
5	79.44 ± 2.97	84.56 ± 3.33	86.78 ± 1.39	79.44 ± 2.53	**92.89 ± 1.67**	86 ± 1.90
7	86.67 + 3.45	87.83 ± 3.32	90.50 ± 2.73	81.5 ± 3.64	**94.33 ± 1.17**	89± 3.06

Table 3. Mean classification accuracy (in percent) over 10 random runs for JAFFE dataset with reduced dimension of 10 using 2DPCA and (2D)^{2}PCA as dimensionality reduction methods.

Training face (#)	2DPCA (256×10)	2DPCA-SVM (256×10)	2DPCA-ProjSVM (256×10)	(2D)^{2}PCA (10×10)	(2D)^{2}PCA-SVM (10×10)	(2D)^{2}PCA-ProjSVM (10×10)
3	91.83 ± 2.30	94.17 ± 1.64	**94.28± 1.34**	91.78 ± 1.94	94.11 ± 2.08	94.06 ± 1.11
5	96.81 ± 2.35	98.05 ± 1.47	97.56 ± 3.38	97.38 ± 2.12	98.06 ± 1.12	**98.13 ± 1.56**
7	98.35 ± 2.37	98.93 ± 1.23	**99.14 ± 1.30**	97.93 ± 1.86	99.09 ± 1.06	98.93 ± 1.31

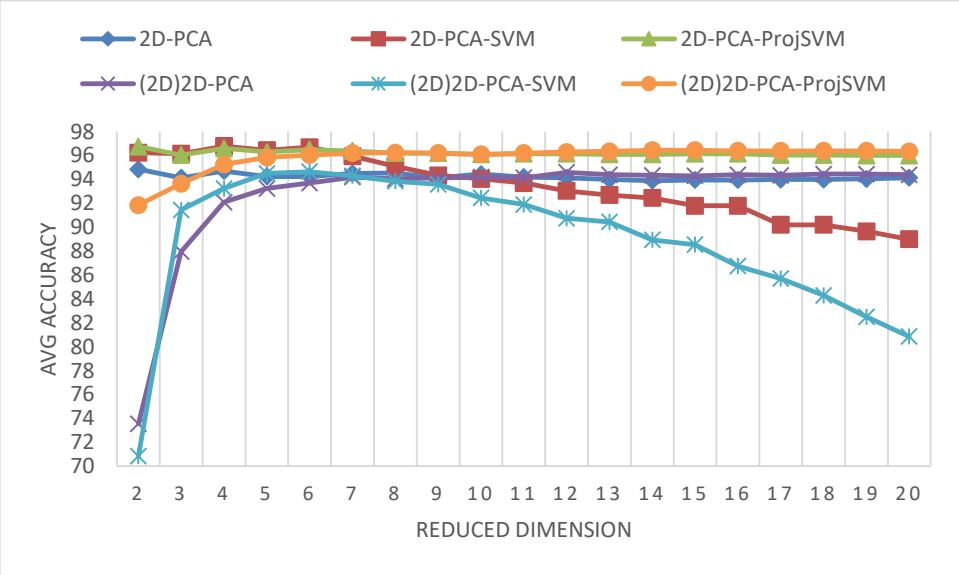

Figure 1. Average accuracy versus reduced dimension for six methods on ORL dataset with PCA-based methods.

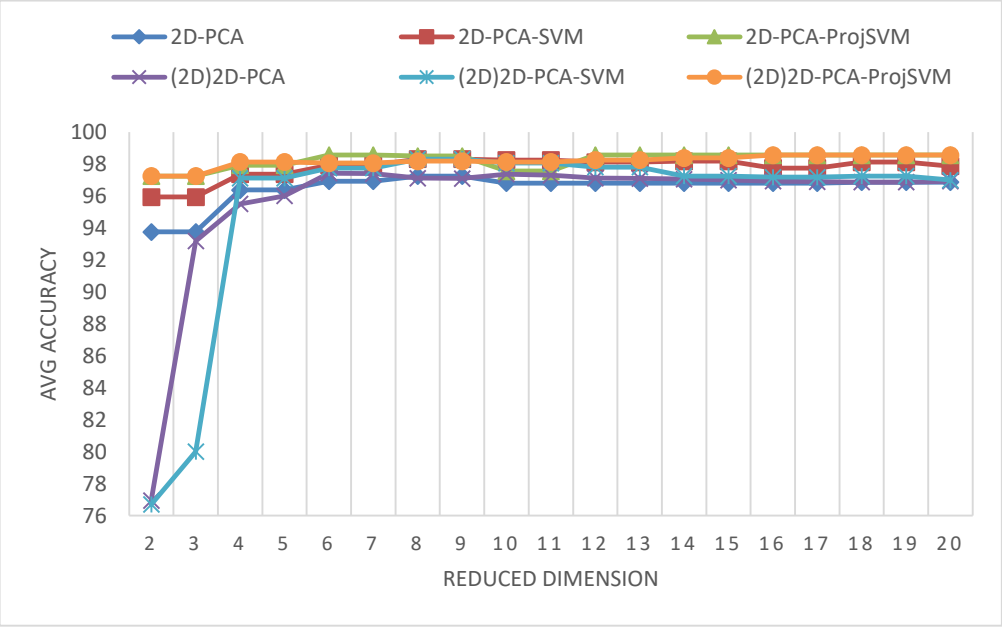

Figure 2. Average accuracy versus reduced dimension for six methods on JAFFE dataset with PCA-based methods.

Figures 3 and 4 show the average accuracy versus the number of dimensions in the reduced space for the above mentioned six methods with 5 training images per subject in 10 random runs on the ORL and JAFFE face datasets, respectively.

In the experiments, we used the two-dimensional (matrix-based) approaches 2DPCA, $(2D)^2$PCA, 2DLDA, and $(2D)^2$LDA, in dimensionality reduction without loss of framework. The main difference between two-directional matrix based approaches, i.e. $(2D)^2$PCA and $(2D)^2$LDA, and the simple matrix-based 2DPCA and 2DLDA approaches is that the latter ones only work in one direction of the face image matrices, while the former ones work in the row and the column directions, simultaneously results in fewer number of coefficients. The experimental results show that $(2D)^2$PCA and $(2D)^2$LDA with smaller feature vectors obtain the same or even higher recognition accuracies compared to the 2DPCA and 2DLDA methods.

Table 4. Mean classification accuracy (in percent) over 10 random runs for ORL dataset with reduced dimension of 10 using 2DLDA and $(2D)^2$LDA as dimensionality reduction methods.

Training face (#)	2DLDA	2DLDA-SVM	2DLDA-ProjSVM	$(2D)^2$LDA	$(2D)^2$LDA-SVM	$(2D)^2$LDA-ProjSVM
3	90.75 ± 1.73	86.32 ± 3.19	**92.32 ± 2.65**	92.12 ± 2.19	82	91.04 ± 1.69
5	95.30 ± 1.47	94.85 ± 1.33	97.25 ± 1.72	95.95 ± 1.23	93.6 ± 2.05	**97.25 ± 1.16**
7	97.33 ± 1.17	98.08 ± 0.97	**98.83 ± 0.86**	97.58 ± 1.49	97 ± 1.05	98.75 ± 0.59

Table 5. Mean classification accuracy (in percent) over 10 random runs for Yale dataset with reduced dimension of 10 using 2DLDA and $(2D)^2$LDA as dimensionality reduction methods.

Training face (#)	2DLDA	2DLDA-SVM	2DLDA-ProjSVM	$(2D)^2$LDA	$(2D)^2$LDA-SVM	$(2D)^2$LDA-ProjSVM
3	86.00 ± 2.83	84.50 ± 3.42	**87.42 ± 3.02**	73.50 ± 5.44	74 ± 4.56	80.92 ± 2.79
5	89.56 ± 3.40	91.44 ± 2.55	91.22 ± 2.31	88.77 ± 2.94	88.89 ± 2.57	**91.44 ± 1.58**
7	92.33 ± 3.78	93.33 ± 2.99	96.17 ± 2.36	93.67 ± 2.92	92.33 ± 1.97	**97 ± 1.36**

Table 6. Mean classification accuracy (in percent) over 10 random runs for JAFFE dataset with reduced dimension of 10 using 2DLDA and $(2D)^2$LDA as dimensionality reduction methods.

Training face (#)	2DLDA	2DLDA-SVM	2DLDA-ProjSVM	$(2D)^2$LDA	$(2D)^2$LDA-SVM	$(2D)^2$LDA-ProjSVM
3	94.11 ± 2.73	93.11 ± 2.18	**94.56 ± 1.49**	92.50 ± 2.16	63.94±7.33	93.94 ± 1.32
5	97.18 ± 2.47	97.44 ± 1.87	**98.38 ± 1.18**	96.93 ± 1.80	79.81 ± 7.97	98.06 ± 1.39
7	98.78 ± 2.17	98.50 ± 1.52	**98.86 ± 1.39**	98.5 ± 1.66	88.86 ± 3.25	**98.86 ± 1.39**

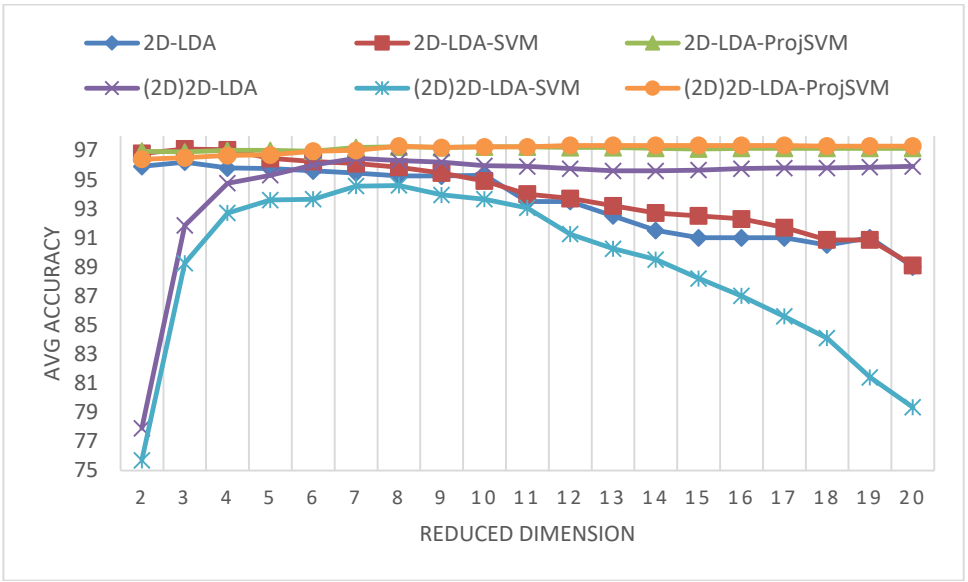

Figure 3. Average accuracy versus reduced dimension for six methods on ORL dataset with LDA-based methods.

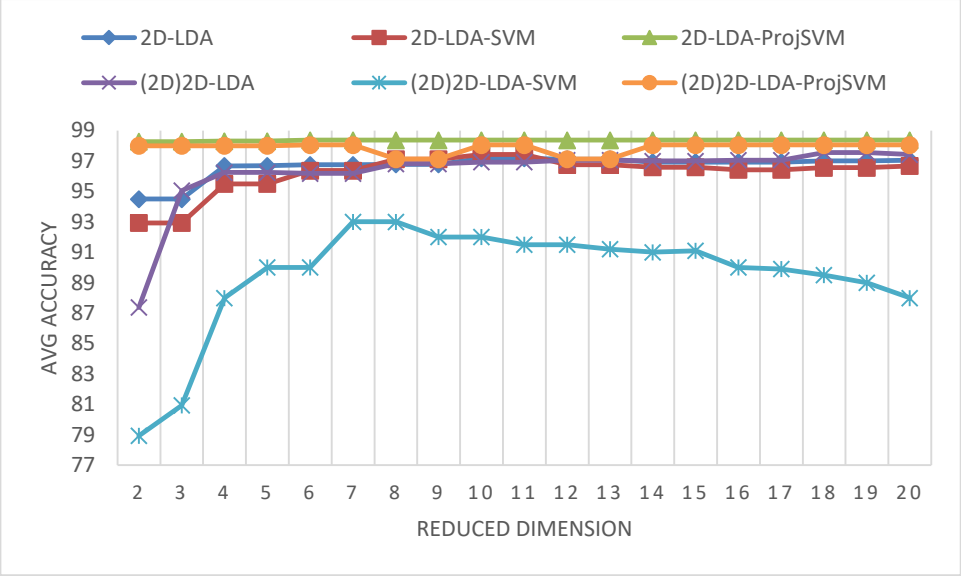

Figure 4. Average accuracy versus reduced dimension for six methods on JAFFE dataset with LDA-based methods.

As observed in the above tables, the accuracies of the proposed methods, compared to the other relevant methods, are better in most of the experiments. According to figures 1, 2, 3, and 4, an increase in the reduced dimension initially increases the accuracy of the methods, in general, and then decreases it. In higher dimensions, where the other methods perform poorly due to the curse of dimensionality, the proposed methods are generally the best ones among the considered techniques.

In addition, the experimental results reported in tables 1-6 indicate that increasing the number of training data, the classification accuracy increases in all the cases, as expected. Finally, the accuracies reported in tables 1 and 2 are generally lower than the corresponding ones in tables 3 and 4. This shows the positive role of the supervision in the dimensionality reduction approach used. The LDA-based methods, due to their supervised nature, result in better performances compared to the PCA-based ones, which are unsupervised dimensionality reduction approaches.

5. Conclusions

In this paper, several methods were proposed to improve the accuracy of face recognition task using the projection penalty idea. In these methods, the feature vectors were first extracted using 2D (matrix-based) extensions of PCA and LDA, and then SVM classifier with projection penalty was used as a predictive model. The projection penalty framework enabled the predictive model to gain from dimensionality reduction techniques without losing relevant information. The experimental results obtained indicated that using these matrix-based linear dimensionality reduction methods in projection penalty framework improved the face recognition performance compared to the classification after dimensionality reduction based on the Euclidean distance or SVM classifier. The superiority of the proposed methods in terms of average accuracies over the other testing methods on three well-known face datasets for various dimensions had been observed through the experiments.

The simple matrix-based 2DPCA and 2DLDA dimensionality reduction methods extracted the features from one dimension of the image matrix and the feature vectors obtained that are smaller than the feature vectors extracted from the whole image matrix. In order to consider both dimensions of the image in the feature extraction step, two-directional 2DPCA and two-directional 2DLDA were used in the proposed framework to perform dimensionality reduction in both the directions of the image matrix generating proper lower-dimensional feature vectors. Good performances for the proposed methods compared to several relevant techniques were observed from the experimental results.

Several new 2D lossless dimensionality reduction methods proposed in this paper are applicable to any image dataset and more generally on any 2D signal samples. In this paper, we chose the face recognition problem as a suitable case study, since dimensionality reduction in this domain was meaningful and had a key role in the final accuracy of the classifiers. The proposed methods can be used in its current form on general image recognition problems such as handwritten

character recognition, and object recognition without any extension or change.

References

[1] Bishop, C. M. (2006). Pattern recognition. Machine Learning, 128.

[2] Van der Maaten, L., Postma, E. & Van Den Herik, H. (2009). Dimensionality reduction: A comparative review. Journal of Machine Learning Research, vol. 10, pp. 1-41.

[3] Adibi, P., & Safabakhsh, R. (2009). Batch linear manifold topographic map with regional dimensionality estimation. Paper presented at the 2009 International Joint Conference on Neural Networks, pp. 63-70.

[4] Tenenbaum, J. B., De Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. Science, vol. 290, no. 5500, pp. 2319-2323.

[5] Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. Science, vol. 290, no. 5500, pp. 2323-2326.

[6] Belkin, M., & Niyogi, P. (2002). Laplacian eigenmaps and spectral techniques for embedding and clustering. In Advances in neural information processing systems, pp. 585-591.

[7] Weinberger, K. Q., & Saul, L. K. (2006). An introduction to nonlinear dimensionality reduction by maximum variance unfolding. Paper presented at the AAAI, vol. 6, pp. 1683-1686.

[8] Fisher, R. A. (1938). The statistical utilization of multiple measurements. Annals of eugenics, vol. 8, no. 4, pp. 376-386.

[9] Scholkopft, B., & Mullert, K.-R. (1999). Fisher discriminant analysis with kernels. Neural networks for signal processing IX, vol. 1, no. 1, pp. 41-48.

[10] Yan, S., Xu, D., Zhang, B., Zhang, H.-J., Yang, Q., & Lin, S. (2007). Graph embedding and extensions: a general framework for dimensionality reduction. IEEE transactions on pattern analysis and machine intelligence, vol. 29, no. 1, pp. 40-51.

[11] Lan, Y.-D., Deng, H., & Chen, T. (2012). Dimensionality reduction based on neighborhood preserving and marginal discriminant embedding. Procedia Engineering, vol. 29, pp. 494-498.

[12] Baghbidi, M., Homayouni, S., Jamshidi, K., & Naghsh-Nilchi, A. R. (2015). Impact of linear dimensionality reduction methods on the performance of anomaly detection algorithms in hyperspectral images. Journal of AI and Data Mining, vol. 3, no. 1, pp. 11–20.

[13] Yang, J., Zhang, D., Frangi, A. F., & Yang, J.-y. (2004). Two-dimensional PCA: a new approach to appearance-based face representation and recognition. IEEE transactions on pattern analysis and machine intelligence, vol. 26, no. 1, pp. 131-137.

[14] Shah, J. H., Sharif, M., Raza, M., & Azeem, A. (2013). A Survey: Linear and Nonlinear PCA Based Face Recognition Techniques. Int. Arab J. Inf. Technol., vol. 10, no. 6, pp. 536-545.

[15] Xu, Y., Zhang, D., Yang, J., & Yang, J.-Y. (2008). An approach for directly extracting features from matrix data and its application in face recognition. Neurocomputing, vol. 71, no. 10, pp. 1857-1865.

[16] Fang, W.-L., Yang, Y.-K., & Pan, J.-K. (2012). A low-computation approach for human face recognition. International Journal of Pattern Recognition and Artificial Intelligence, vol. 26, no. 06, p. 1256015.

[17] Li, M., & Yuan, B. (2005). 2D-LDA: A statistical linear discriminant analysis for image matrix. Pattern Recognition Letters, vol. 26, no. 5, pp. 527-532.

[18] Herwindiati, D. E., Isa, S. M., & Hendryli, J. (2014). Performance of robust two-dimensional principal component for classification. Paper presented at the Advanced Computer Science and Information Systems (ICACSIS), 2014 International Conference on, pp. 434-440.

[19] Zhang, D. & Zhou, Z. H. (2005). $(2D)^2$PCA: 2-Directional 2-Dimensional PCA for Efficient Face Representation and Recognition, "journal of Neurocomputing, vol. 69, pp. 224-231.

[20] Noushath, S., Kumar, G. H., & Shivakumara, P. (2006). $(2D)^2$LDA: An efficient approach for face recognition. Pattern Recognition, vol. 39, no. 7, pp. 1396-1400.

[21] Wang, D., Shen, H., & Truong, Y. (2016). Efficient dimension reduction for high-dimensional matrix-valued data. Neurocomputing, vol. 190, pp. 25-34.

[22] Zhang, Y., & Schneider, J. G. (2010). Projection penalties: dimension reduction without loss. Paper presented at the Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 1223-1230.

[23] Ahmadkhani, S., & Adibi, P. (2016). Face recognition using supervised probabilistic principal component analysis mixture model in dimensionality reduction without loss framework. IET Computer Vision, vol. 10, no. 3, pp. 193-201.

[24] Guo, Z., Wang, H., & Liu, Q. (2012). $(2D)^2$LDALPP: A Novel Approach to Face Recognition. International Journal of Advanced Robotic Systems, vol. 9, no. 5, p.221.

[25]http://vismod.media.mit.edu/vismod/classes/mas622-00/datasets/