A Simple and Efficient Supervised Method for Spatially Weighted PCA in Face Image Analysis

Carlos E. Thomaz
Department of Electrical Engineering, FEI
São Bernardo do Campo, São Paulo, Brazil

Gilson A. Giraldi
National Laboratory for Scientific Computing
Petrópolis, Rio de Janeiro, Brazil

Joaquim F. P. da Costa
Department of Applied Mathematics, University of Porto
Porto, Portugal

Duncan F. Gillies
Department of Computing, Imperial College London
London, UK

August 10, 2010

Abstract

Principal Component Analysis (PCA) is an example of a successful unsupervised statistical dimensionality reduction method, especially in small sample size problems. Despite the well-known attractive properties of PCA, the traditional approach does not incorporate prior information extracted from a specific domain knowledge. The development of techniques that bring together dimensionality reduction and prior knowledge can be performed in the framework of supervised learning methods, like Fisher Discriminant Analysis. Semi-supervised methods can also be applied if only a small number of labeled samples is available. In this paper, we propose a simple and efficient supervised method that allows PCA to incorporate explicitly domain knowledge and generates an embedding space that inherits its optimality properties for dimensionality reduction. The method relies on discriminant weights given by separating hyperplanes to generate the spatially weighted PCA. Several experiments using 2D frontal face images and different data sets have been carried out to illustrate the usefulness of the method for dimensionality reduction, classification and interpretation of face images.

1 Introduction

The goal of dimensionality reduction is to define a compact representation of the original data by embedding high-dimensional data samples into a lower dimensional space that preserves most

of the intrinsic information contained in the data [16, 13, 24]. The obtained representation can be used for various succeeding tasks such as data visualization, understanding, classification and mining [11, 30, 12, 6].

The Principal Component Analysis (PCA) [20, 15] is the most known multivariate statistical linear method for dimensionality reduction and has been applied successfully in a number of supervised problems, such as face recognition [22, 28, 26], as a pre-processing step to reduce the computational costs, mitigate the curse of dimensionality [8] and improve the classification performance. Despite the well-known attractive properties of PCA, it remains challenging the issue of how to include prior information in the PCA formulation and not discard important discriminant information related to the principal components with the smallest eigenvalues.

The development of techniques that bring together dimensionality reduction and prior knowledge can be performed in the framework of supervised learning approaches, like the discriminant feature extraction methods based on the Fisher's criterion [5, 8]. However, a critical limitation of such methods for dimensionality reduction is that the dimension of the resulting embedding space should be less than the number of classes. Thus, for instance, when there are two sample groups to separate, supervised learning approaches can identify only one meaningful discriminant direction and consequently additional information important to characterize the sample group differences may be lost [2, 33, 9].

In fact, when only a small number of labeled samples are available, multivariate supervised dimensionality reduction methods tend to perform poorly due to overfitting [10]. In [25], such question has been addressed by proposing a semi-supervised dimensionality reduction method that smoothly bridges the gap between supervised and non-supervised approaches, controlling the reliance on the global structure of unlabeled samples and the information brought by labeled samples. Other works on semi-supervised dimensionality reduction methods have also been proposed recently to incorporate the structure of original high-dimensional data and pairwise constraints specified by users, using labeled and unlabeled data simultaneously [31, 1, 17]. However, a common issue to all semi-supervised learning techniques is how to optimize the regularization parameters necessary to blend supervised and non-supervised information often represented by local and global scatter matrices.

As stated recently by Skocaj et al. [23], a reliable and robust subspace representation of high-dimensional data should essentially enable a selective treatment of the individual attributes that compose the patterns of interest. For instance, in visual learning and image recognition, we can understand the pixels of n-dimensional image vectors as attributes of the patterns of interest, which are more (or less) informative depending on the corresponding spatial weights. This requirement is not satisfactorily met by the standard PCA approach [23] that can be extended to a weighted version by introducing a spatial weighting value for each pixel in the image [23]. Such approach can be used for time series analysis as well. In this case, we can define different temporal weights for individual observations rather than individual attributes, determining, for instance, that the most recent samples or data may have a larger influence on the estimation of the principal subspace than others [3]. Therefore, in the last years, weighted PCA techniques have been proposed in the context of eigenspace learning [23], motion data analysis [7], time-series compression [3] and palmprint identification [32] in order to obtain a consistence subspace representation of the original data in the presence of noise, outliers and missing data. Despite of the success of these methods, a key remaining question for the weighted PCA methods in general is how to automatically compute the corresponding spatial (or temporal) weights that could take advantage of some labeled information available without jeopardizing the PCA unsupervised nature and its inherent straightforward and simple calculation.

In this paper, we address this issue through supervised learning techniques. We propose a supervised weighted PCA that incorporates domain knowledge and generates an embedding space (with the same dimension of the original one) that preserves the optimality properties of dimensionality reduction and interpretability of the standard PCA. As a supervised method, we need a set of training samples with some labeled data as the input. Then, a separating hyperplane based on a discriminant criterion is computed and the obtained discriminant weights are used to generate the spatially weighted PCA subspace. In this sense, the weights incorporate the prior knowledge extracted from the labeled data and can be systematically computed through the hyperplane directions. The approach is in fact a simple and natural way of addressing the well-known trade-off between achieving best reconstruction (unsupervised goal) and being most efficient for making predictions (supervised goal) [10] when performing data feature extraction and dimensionality reduction whether some labeled data are available.

The paper is organized as follows. Next, in section 2, we briefly review the standard PCA, its calculation and inherent limitation of providing distinct importance (or weight) to the original variables when some labeled information are available. Then, in section 3, we explain our idea of incorporating domain knowledge in a spatially weighted PCA, which allows an automatic selective treatment of the variables that compose the patterns of interest. Then, section 4 describes all the experiments carried out in this study, as well as the three face databases used to evaluate the effectiveness of the spatially weighted principal components. In section 5, we discuss geometrically the main difference between the spatially weighted principal components and discriminant principal components, recently proposed [**?**]. Finally, in section 6, we conclude the paper, summarizing its main contributions and results.

2 Principal Component Analysis

Principal Component Analysis (**PCA**) [15], also called the Karhunen-Loeve transformation, is a multivariate statistical non-supervised method for data interpretation and dimensionality reduction concerned with explaining the covariance structure of a set of n original variables through a small number of m linear and orthonormal combinations of these variables, where $m \leq n$.

Let an $N \times n$ training set matrix X be composed of N input samples (or patterns of interest, such as face images) with n variables (or attributes, such as pixels), that is, $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}^T$. This means that each column of matrix X represents the values of a particular variable observed all over the N samples. Let this data matrix X have covariance matrix

$$S = \frac{1}{(N-1)} \sum_{i=1}^{N} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T, \tag{1}$$

where $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{in}]^T$ and $\bar{\mathbf{x}}$ is the grand mean vector of X given by

$$\bar{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n). \tag{2}$$

Let the covariance matrix S have respectively P and Λ eigenvector and eigenvalue matrices, that is,

$$P^T S P = \Lambda. \tag{3}$$

It is a proven result that the set of m ($m \leq n$) eigenvectors of S, which corresponds to the m largest eigenvalues, minimizes the mean square reconstruction error over all choices of m orthonormal basis vectors [8]. Such a set of eigenvectors that defines a new uncorrelated coordinate system for the training set matrix X is known as the principal components.

To note explicitly the spatial association between the j^{th} and k^{th} variables, we can rewrite the sample covariance matrix S described in equation (1) in order to indicate the position of each variable in the N samples. When n variables are observed on each sample, the sample variation can be described by the following sample variance-covariance equation [14]

$$S = \{s_{jk}\} = \left\{ \frac{1}{(N-1)} \sum_{i=1}^{N} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) \right\}, \tag{4}$$

for $j = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, n$. The covariance s_{jk} between the j^{th} and k^{th} variables reduces to the sample variance when $j = k$, $s_{jk} = s_{kj}$ for all j and k, and the covariance matrix S contains n variances and $\frac{1}{2}n(n-1)$ potentially different covariances [14].

It is clear from equation (4) that the variable deviations from the mean have the same importance in the standard sample covariance matrix S formulation. In other words, all the n variables are equally weighted. However, there are situations where this should not be the case, particularly in image recognition problems where some parts of the images might be more informative for separating sample groups than others. Moreover, since the standard PCA explains the covariance structure of all the data using only unsupervised information its most expressive components [26], that is, the first principal components with the largest eigenvalues, are not necessarily the most discriminant ones when some labeled information are available.

In the next section, we will define a weighted sample correlation equation that will mitigate these well-known issues inherent to the standard PCA and will enable an automatic selective treatment of the variables that compose the patterns of interest depending on their corresponding spatial discriminant weights.

3 A Supervised Spatially Weighted PCA

In this section, we describe the supervised method proposed to incorporate domain knowledge in a spatially weighted PCA, allowing an automatic selective treatment of the variables that compose the patterns of interest.

3.1 Definition

Let us start by describing the well-known Pearson's sample correlation coefficient between the j^{th} and k^{th} variables, defined as follows [14]:

$$\begin{aligned} r_{jk} &= \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}} \\ &= \frac{\sum_{i=1}^{N}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^{N}(x_{ij} - \bar{x}_j)^2}\sqrt{\sum_{i=1}^{N}(x_{ik} - \bar{x}_k)^2}}, \end{aligned} \tag{5}$$

for $j = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, n$. It is important to note that $r_{jk} = r_{kj}$ for all j and k. From equation (5), it is clear that the sample correlation coefficient is a normalized version of the sample covariance, where the product of the square roots of the sample variances, known as the sample standard deviations, provides the spatial normalization of the sum of the variable deviations from the mean [14]. In other words, r_{jk} is a measure of the linear association between two variables that does not allow that variables with larger variance dominate the corresponding deviations from the mean.

In our model, we want to give higher importance to the variables with higher discriminant weights, but variables that vary most are not necessarily the ones that allow best separation between the sample groups. Therefore, we need to define a measure of association between variables, based on the Pearson's sample correlation coefficient, which uses the notion of spatial weights and is more or less dominant depending on the values of each spatial weight. Following equation (5), we can define the weighted sample correlation r^*_{jk} between the j^{th} and k^{th} variables by

$$\begin{aligned} r^*_{jk} &= \frac{(\sqrt{w_j}\sqrt{w_k})s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}} \\ &= \frac{\sum_{i=1}^{N}\sqrt{w_j}(x_{ij}-\bar{x}_j)\sqrt{w_k}(x_{ik}-\bar{x}_k)}{\sqrt{\sum_{i=1}^{N}(x_{ij}-\bar{x}_j)^2}\sqrt{\sum_{i=1}^{N}(x_{ik}-\bar{x}_k)^2}}, \end{aligned} \tag{6}$$

for $j = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, n$. The spatial weighting vector

$$\mathbf{w} = [w_1, w_2, \ldots, w_n]^T \tag{7}$$

is such that $w_j \geq 0$ and $\sum_{j=1}^{n} w_j = 1$. So, when n variables are observed on N samples, the weighted sample correlation matrix R^* can be described by

$$R^* = \left\{r^*_{jk}\right\} = \left\{ \frac{\sum_{i=1}^{N}\sqrt{w_j}(x_{ij}-\bar{x}_j)\sqrt{w_k}(x_{ik}-\bar{x}_k)}{\sqrt{\sum_{i=1}^{N}(x_{ij}-\bar{x}_j)^2}\sqrt{\sum_{i=1}^{N}(x_{ik}-\bar{x}_k)^2}} \right\}. \tag{8}$$

for $j = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, n$. Analogously to equation (4), the sample correlation r^*_{jk} between the j^{th} and k^{th} variables is equal to w_j when $j = k$, $r^*_{jk} = r^*_{kj}$ for all j and k, and the weighted correlation matrix R^* is a nxn symmetric matrix.

Let the weighted correlation matrix R^* have respectively P^* and Λ^* eigenvector and eigenvalue matrices, that is,

$$P^{*T} R^* P^* = \Lambda^*. \tag{9}$$

The set of m ($m \leq n$) eigenvectors of R^*, that is, $P^* = [\mathbf{p}^*_1, \mathbf{p}^*_2, \ldots, \mathbf{p}^*_m]$, which corresponds to the m largest eigenvalues, defines a new orthonormal coordinate system for the training set matrix X and is called here as the *spatially weighted principal components*.

The remaining question now is: how to define spatial weights w_j that incorporate the prior knowledge extracted from the labeled data and can be systematically computed through the supervised information available? In this paper, we address this task through spatial weights obtained as the output of a linear learning process for separating tasks, as described in the next subsection.

3.2 The Spatial Discriminant Weights

We propose the idea of using the discriminant weights given by statistical separating hyperplanes as the spatial weights of the weighted sample correlation matrix defined in equation (8). Such approach is not restricted to any particular probability density function of the sample groups because it can be based on either a parametric or non-parametric separating hyperplane approach.

As supervised methods, the models need some labeled data of N pairs

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N), \tag{10}$$

where $\mathbf{x}_i \in \Re^n$ denote the i^{th} training observations and y_i are scalars that correspond to the classification labels. For simplicity and without loss of generality, we concentrate on two-class problems, that is, $y_i \in \{-1, 1\}$.

Let us start with the parametric spatial weights given by Linear Discriminant Analysis (LDA) based approaches [5, 8]. The standard LDA solution is a spectral matrix analysis of the data and is based on the assumption that each class can be represented by its distribution of data, that is, the corresponding mean vector (or class prototype) and covariance matrix (or spread of the sample group) [12]. LDA depends on all of the data, even points far away from the separating hyperplane and its main objective is to find a projection vector $\mathbf{w}_{lda}$ that maximizes the Fisher's criterion [8]:

$$\mathbf{w}_{lda} = \arg\max_{w} \frac{\left|\mathbf{w}^T S_b \mathbf{w}\right|}{\left|\mathbf{w}^T S_w \mathbf{w}\right|}. \tag{11}$$

The S_b and S_w matrices are the between-class and within-class scatter matrices defined as

$$S_b = \sum_{i=1}^{g} N_i (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^T, \tag{12}$$

$$\begin{aligned} S_w &= \sum_{i=1}^{g} (N_i - 1) S_i \\ &= \sum_{i=1}^{g} \sum_{j=1}^{N_i} (\mathbf{x}_{i,j} - \bar{\mathbf{x}}_i)(\mathbf{x}_{i,j} - \bar{\mathbf{x}}_i)^T, \end{aligned} \tag{13}$$

where $\mathbf{x}_{i,j}$ is the n-dimensional pattern (or observation) j from class i, N_i is the number of training patterns from class i, and g is the total number of classes or groups, that is, $g = 2$ in a two-class problem. The vector $\bar{\mathbf{x}}_i$ and matrix S_i are respectively the unbiased sample mean and sample covariance matrix of class i [8]. The Fisher's criterion is maximized when the projection vector $\mathbf{w}_{lda}$ is the leading eigenvector of $S_w^{-1} S_b$, assuming that S_w is invertible.

However, in small sample size problems, S_w is either singular or mathematically unstable and the standard LDA cannot be used to the classification task. To avoid these critical issues, we have calculated the leading eigenvector $\mathbf{w}_{lda}$ by using two different approaches. The first approach, based on the Zhu and Martinez method [36, 34, 35], replaces S_w with the nxn identity matrix and $\mathbf{w}_{lda}$ becomes simply the leading eigenvector of S_b. The other, based on the Maximum uncertainty Linear Discriminant Analysis (MLDA) proposed by Thomaz et al. [27], considers the issue of regularizing the S_w estimate with a multiple of the identity matrix. In both approaches, when $g > 2$, the number of leading discriminant eigenvectors with non-zero eigenvalues is limited by the rank of S_b and consequently is equal to $\min\{n, g - 1\}$.

To allow the investigation of spatial discriminant weights determined by separating hyperplanes that do not make any assumption on the distribution of the data, we have used the Support Vector Machine method [29] based on the risk-minimization approach. The primary purpose of SVM is to maximize the width of the margin between two distinct sample classes [29]. Given a training set as described in the formulation (10), the SVM method seeks to find the hyperplane defined by

$$f(\mathbf{x}) = (\mathbf{x} \cdot \mathbf{w}) + b = 0, \tag{14}$$

which separates positive and negative observations with the maximum margin. It can be shown that the solution vector $\mathbf{w}_{svm}$ is defined in terms of a linear combination of the training observations, that is,

$$\mathbf{w}_{svm} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i, \tag{15}$$

where α_i are non-negative coefficients obtained by solving a quadratic optimization problem with linear inequality constraints. Those training observations $\mathbf{x}_i$ with non-zero α_i lie on the boundary of the margin and are called support vectors [29]. Thus, the SVM separating hyperplane solution is based essentially on the observations that lie close to the opposite class, that is, on the observations that most count for classification [12].

We have focused here on two parametric separating hyperplanes based on spectral matrix analyzes and one non-parametric separating hyperplane based on the risk minimization criterion. However, any other separating hyperplane could be used as long as the supervised information provided by the labeled data is incorporated on the separating hyperplane calculation.

3.3 The Step-by-Step Algorithm

The main steps for calculating the spatially weighted principal components $P^* = [\mathbf{p}_1^*, \mathbf{p}_2^*, \ldots, \mathbf{p}_m^*]$ of an $N \times n$ training set matrix X composed of N input samples with n variables can be described as follows:

1. Calculate the spatial weighting vector $\mathbf{w} = [w_1, w_2, \ldots, w_n]^T$ using some labeled data and a separating hyperplane method, as described in the previous sub-section;

2. Normalize $\mathbf{w}$ such that $w_j \geq 0$ and $\sum_{j=1}^{n} w_j = 1$, that is replace w_j with $\frac{|w_j|}{\sum_{j=1}^{n} |w_j|}$;

3. Standardize all the n variables of the data matrix X such that the new variables have $\bar{x}_j = 0$ and $s_j = s_{jj} = 1$, for $j = 1, 2, \ldots, n$. In other words, calculate the grand mean vector

$$\bar{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n)$$

and the variance vector

$$\mathbf{s} = \frac{1}{N} \sum_{i=1}^{N} (x_{ij} - \bar{x}_j)^2 = (s_1, s_2, \ldots, s_n)$$

and replace x_{ij} with z_{ij} given by

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{s_j}}$$

for $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, n$;

4. Spatially weigh up all the standardized z_{ij} variables using the normalized weighting vector $\mathbf{w}$ calculated in step 2, that is

$$z^*_{ij} = z_{ij}\sqrt{w_j};$$

5. The spatially weighted principal components P^* are then the eigenvectors corresponding to the m largest eigenvalues of $(Z^*)^T Z^*$, where $Z^* = \{\mathbf{z}^*_1, \mathbf{z}^*_2, \ldots, \mathbf{z}^*_N\}^T$.

In small sample size problems, where $N \ll n$, standard PCA can be used previously to the first step of the algorithm for dimensionality reduction to alleviate the computational costs of the further learning method used. In this case, the spatial weighting vector $\mathbf{w}$ in the original n-dimensional space can be obtained by the expression $\mathbf{w} = P \cdot \mathbf{w}_{pca}$, where P is composed of all the standard principal components with non-zero eigenvalues and $\mathbf{w}_{pca}$ is the spatial weighting vector calculated on this standard PCA transformed subspace.

4 Experimental results

We have divided our experimental results into three parts. Firstly, we have carried out some face and facial expression image analyzes to understand and visualize the spatial weights found by the separating hyper-planes. Then, in the second part, we have investigated the usefulness of the weighted principal components on recognizing samples compared to the standard unsupervised PCA. Finally, in the last part, we have analyzed the effectiveness of the weighted principal components on reconstructing samples compared to the standard unsupervised PCA.

The following two-group separation tasks have been performed using frontal face images: (a) Gender experiments (female versus male samples); (b) Facial expression experiments (non-smiling versus smiling, anger versus disgust, happiness versus sadness and fear versus surprise samples). The goal of the gender experiment is to evaluate the weighting method on a discriminant task where the differences between the groups are evident. The facial expression experiment poses an alternative analysis where there are subtle differences between the groups.

In all experiments, the total number of training examples N is limited and significantly less than the dimension of the feature space, that is, $N \ll n$ (*small sample size problem*) [8]. To address this problem for the Fisher's criterion, we have used a regularized version of the LDA approach called MLDA [27], as describe in the subsection 3.2.

4.1 Understanding and visualizing the spatial weights

In these experiments, we have used frontal and pre-aligned images of a face database maintained by the Department of Electrical Engineering of FEI[1], São Paulo, Brazil. In this data set the number of subjects is equal to 200 (100 men and 100 women) and each subject has two frontal images (one with a neutral or non-smiling expression and the other with a smiling facial expression), a total of 400 images to perform the experiments. All faces are mainly represented by subjects between 19 and 40 years old with distinct appearance, hairstyle and adorns.

[1]This database is publicly available on http://www.fei.edu.br/∼cet/facedatabase.html

As the average face image is an n-dimensional point that retains all common features from the training sets, we could use this image to understand and visualize the spatial weights found by the separating hyper-planes. Figure 1 illustrates the spatial distribution of the discriminant weights extracted by each separating hyper-plane superimposed on the average face image. Face regions contained within the colored areas and closer to the spectrum of yellow and white show pixels of relatively larger discriminant weights (in absolute values).

We can see clearly that by exploring the discriminant information captured by the separating hyper-planes we are able to identify features that most differ between the sample groups [9]. Also, such discriminant information varies depending on the separating hyper-plane used. For instance, in these experiments where the images are well-framed, the discriminant weights extracted by the Zhu and Martinez separating hyper-plane seem to be more informative and less prone to overfitting for characterizing group-differences than the MLDA and SVM ones.

Figure 2 shows the histograms of the spatial weights illustrated on Figure 1 for all the separating hyper-planes on the gender and expression discriminant tasks. As we should expect, since the sample group differences related to the facial expression discriminant task are subtler and more localized than the gender one, the number of pixels of the expression task at the 2D positions (i, j) considered as non-discriminant, that is, $w_{i,j} \simeq 0$, is much higher than the corresponding ones of the gender experiments. For instance, with the exception of the SVM separating hyper-plane, both Zhu and Martinez and MLDA methods have more than 4000 pixels that would be considered non-discriminant ones in the expression experiments, which represent approximately twice the corresponding number of non-discriminant pixels for the gender experiments. In fact, based on this histogram information we could assess the discriminant significance of each pixel and select only the top ranking ones, increasing significantly the reduction of the dimensionality of the original space performed by PCA for specific separating tasks.

4.2 Recognition rates of the weighted components

In the second part of the experimental results, we have investigated the usefulness of the weighted principal components on recognizing samples. Besides the FEI face database described in the previous subsection, we have used two other data sets to evaluate the classification performance of the spatially weighted principal components: the well-known FERET [21] and Japanese Female Facial Expression (JAFFE) [18] databases. In the FERET database, we have considered 200 subjects (107 men and 93 women) and each subject has two frontal images (one with a neutral or non-smiling expression and the other with a smiling facial expression), providing a total of 400 images to perform the gender and expression experiments. The JAFFE database is a facial expression data set composed of 193 images of expressions posed by nine Japanese subjects, all women. Each person posed three or four examples of each of six fundamental facial expression: anger, disgust, fear, happiness, sadness and surprise. This database has at least 29 images for each fundamental facial expression and we have performed the following two-group classification tasks: happiness versus sadness, fear versus surprise and anger versus disgust.

We have adopted the 10-fold cross validation method to evaluate the classification performance of the standard and weighted PCA methods. Throughout all the classification experiments, we have assumed that the prior probabilities and misclassification costs are equal for both groups. On the PCA subspace, the mean of each class i has been calculated from the corresponding training images and the Mahalanobis distance from each class mean $\overline{x}_i$ has been

Figure 1: Understanding and visualizing the spatial weights found by the following separating hyper-planes (from top to bottom): Zhu and Martinez method, MLDA and SVM. On the left and right columns, we can see the discriminant weights of the corresponding separating hyper-planes related to the gender and facial expression experiments, respectively, using the FEI database.

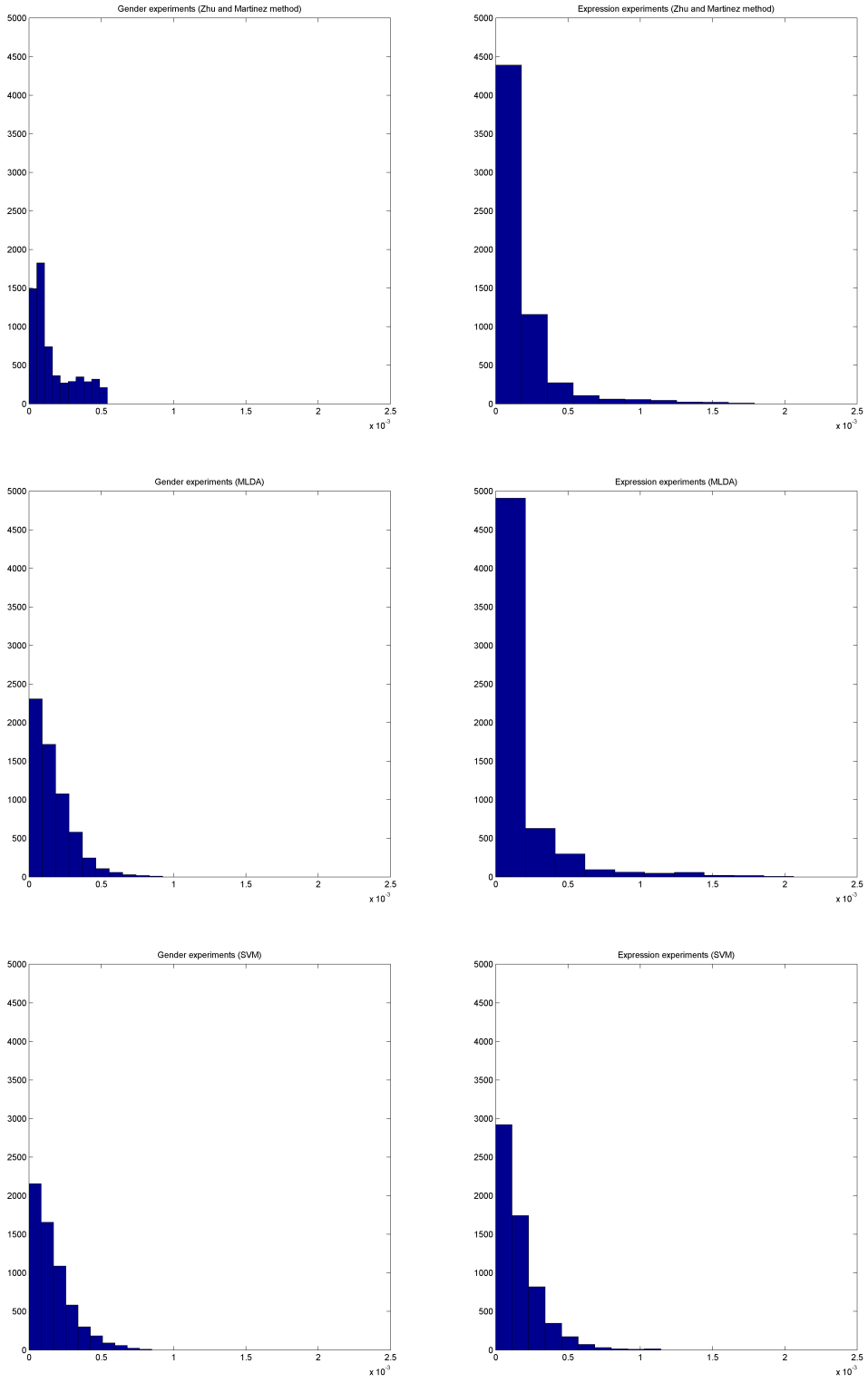

Figure 2: Histogram of the spatial weights found by the following separating hyper-planes (from top to bottom): Zhu and Martinez method, MLDA and SVM. On the left and right columns, we can see the discriminant weights of the corresponding separating hyper-planes related to the gender and facial expression experiments, respectively, using the FEI database.

used to assign a test observation x_t to either the male or female groups in the gender experiment, or to either the different facial expressions in the expression experiment. That is, we have assigned $\mathbf{x}_t$ to class i that minimizes:

$$d_i(\mathbf{x}_t) = \sum_{j=1}^{k} \frac{1}{\lambda_j}(x_{tj} - \overline{x}_{ij})^2, \tag{16}$$

where λ_j is the corresponding eigenvalue and k is the number of principal components retained. In all recognition experiments, we have considered different number of principal components to calculate the recognition rates of the methods implemented.

Figure 3 shows the average recognition rate of the 10-fold cross validation of the gender experiments using the FEI and FERET databases and different number of principal components selected by the corresponding largest eigenvalues. It is possible to see that even in such experiments where the differences between the sample groups are not subtle, the use of prior information given by labeled samples improves the discriminant power of the principal components, allowing similar or higher average recognition rates with the same number of components. For instance, in the gender experiments using the FEI database, all the spatially weighted PCA methods consistently outperform the standard PCA when the number of principal components retained has been higher than 10, that is, when $k \geq 10$. In the gender experiments using the FERET database, which is composed of frontal face images not as well aligned as in the FEI database, the superiority of the spatially weighted PCA is not so clear, but still it is possible to see a better classification performance than the standard PCA when using few principal components, that is, when $5 \leq k < 40$.

The importance of allowing a discriminant treatment of individual pixels and, consequently, minimizing the potential problem of discarding information related to subtle group differences on the first components of the standard PCA can be seen in Figure 4. In both FEI and FERET face databases, the average recognition rates of the spatially weighted principal components are much higher than the standard ones when the original dimensionality of the data is considerably reduced. For example, when using only $k = 5$ spatially weighted principal components, it is possible to achieve an average recognition rate of approximately 92% compared to 55% with the standard PCA. A significant improvement in classification performance is illustrated as well in the expression experiments using the FERET face database, where the spatially weighted and the standard principal components have achieved respectively approximately 70% and 55% with $k = 40$ components, for instance.

An analogous improvement in the classification performance of the weighted principal components compared to the standard ones can be seen in the JAFFE database experiments, shown in Figures 5-7. In these pair-wisely facial expression results where differences between the sample groups can be minor, we can see that the weighted principal components achieved the highest recognition rates in all experiments, building a more efficient feature representation of the data not only in terms of making predictions but also in terms of reducing the dimensionality of the original images.

4.3 Reconstruction based on the weighted components

Another important point of building a reduced feature representation of the original data, especially in the image domain, is related to the issue of achieving good reconstruction or data compression on an unsupervised manner [10] without losing the discriminant information for

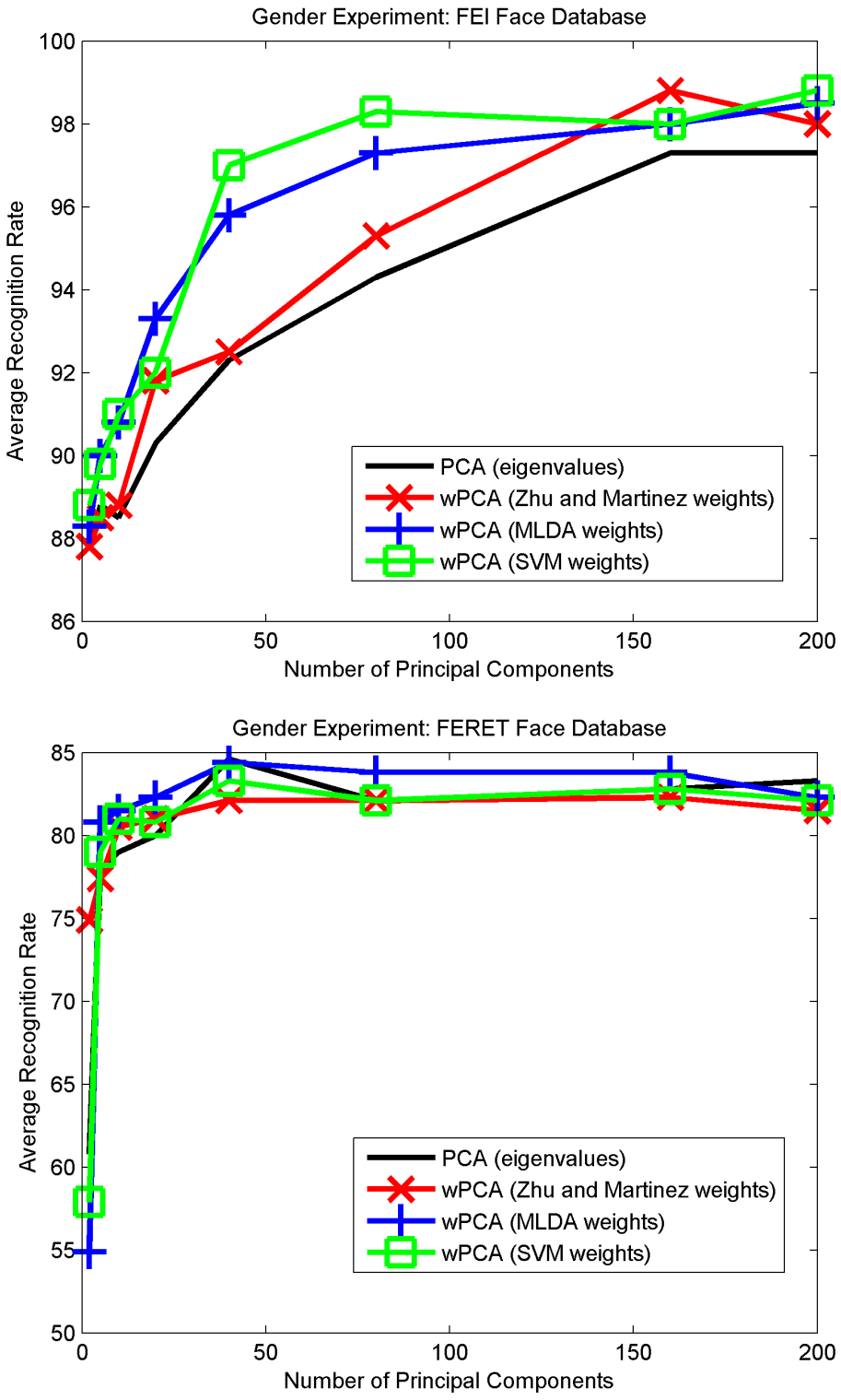

Figure 3: Gender recognition performance of the spatially weighted PCA (wPCA) compared to the standard PCA using the FEI (top) and FERET (bottom) databases. All the principal components retained have been selected by their corresponding largest eigenvalues.

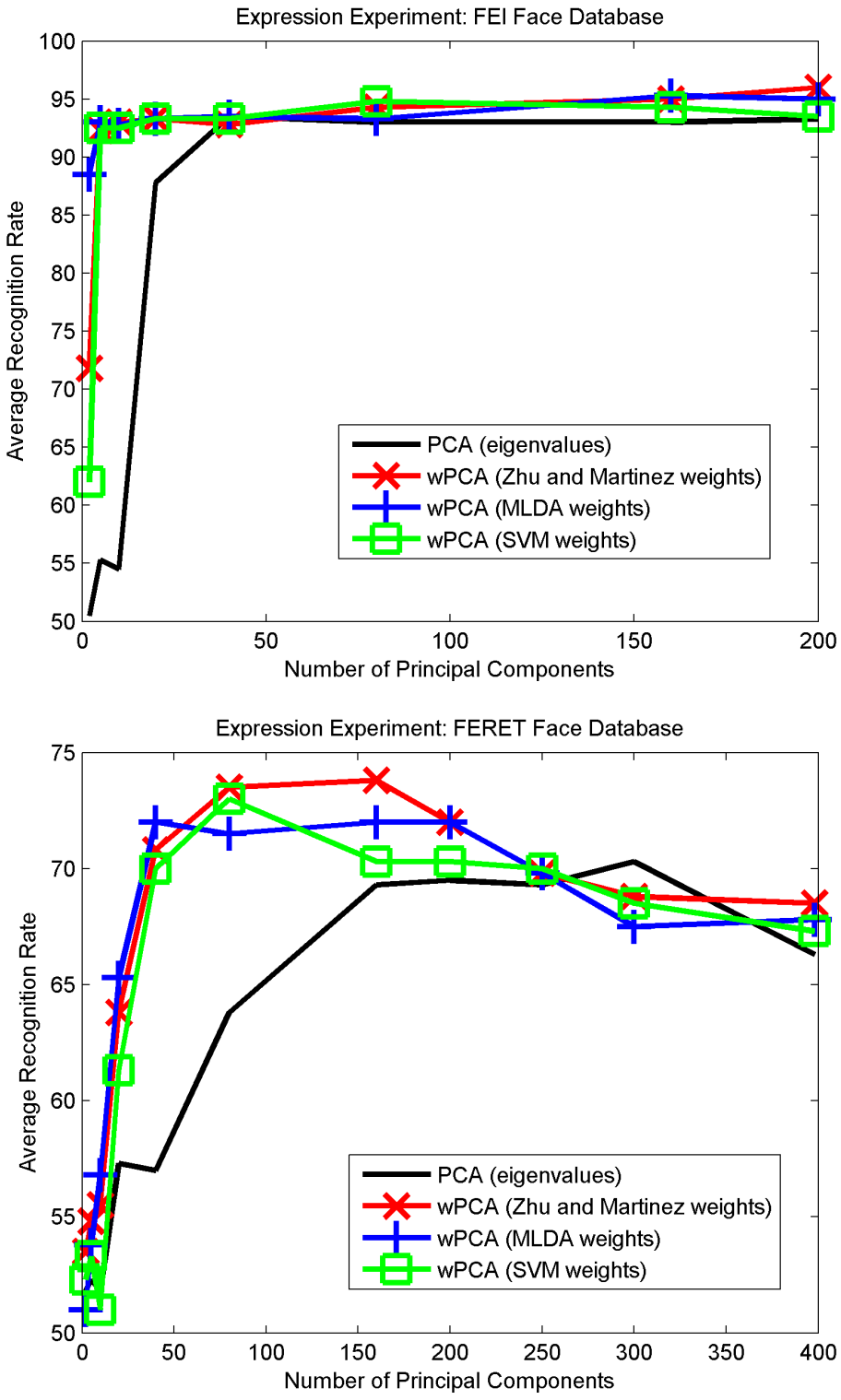

Figure 4: Expression recognition performance of the spatially weighted PCA (wPCA) compared to the standard PCA using the FEI (top) and FERET (bottom) databases. All the principal components retained have been selected by their corresponding largest eigenvalues.

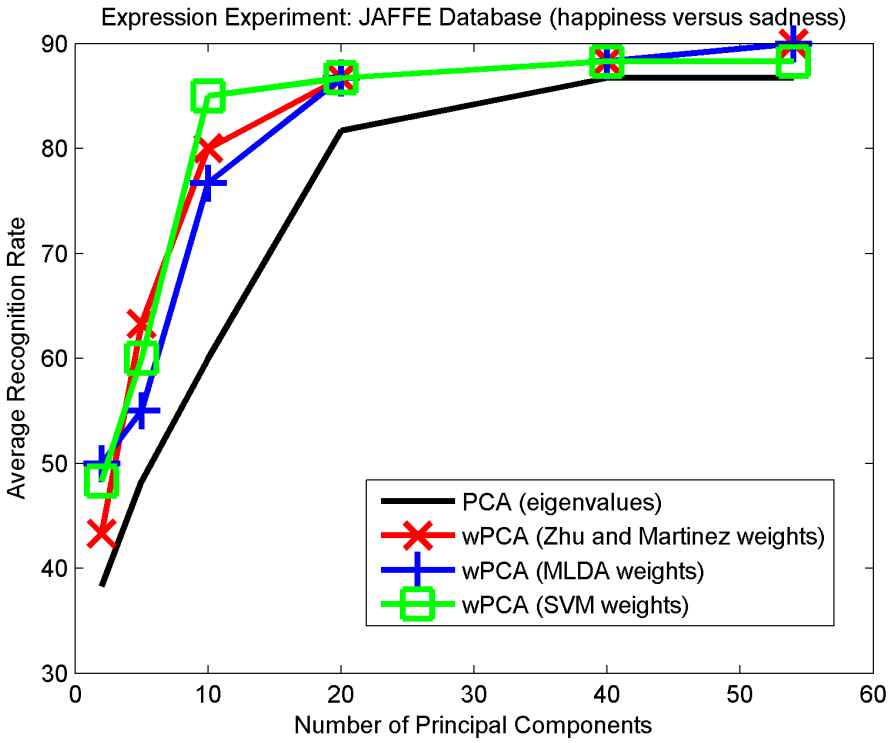

Figure 5: Happiness versus sadness facial expression recognition performance of the spatially weighted PCA (wPCA) compared to the standard PCA using the JAFFE database. All the principal components retained have been selected by their corresponding largest eigenvalues.

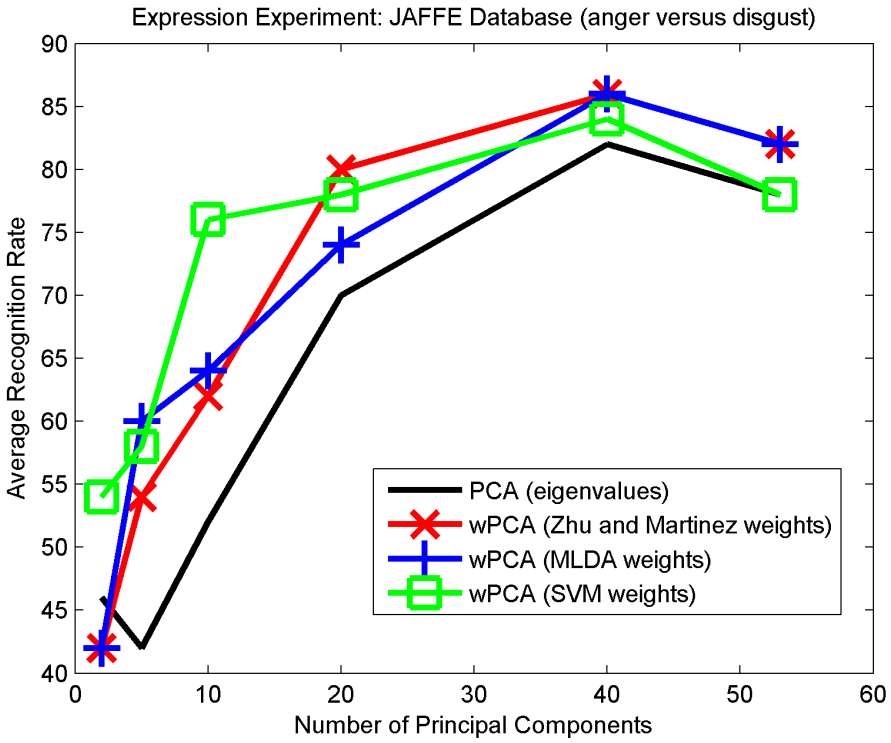

Figure 6: Anger versus disgust facial expression recognition performance of the spatially weighted PCA (wPCA) compared to the standard PCA using the JAFFE database. All the principal components retained have been selected by their corresponding largest eigenvalues.

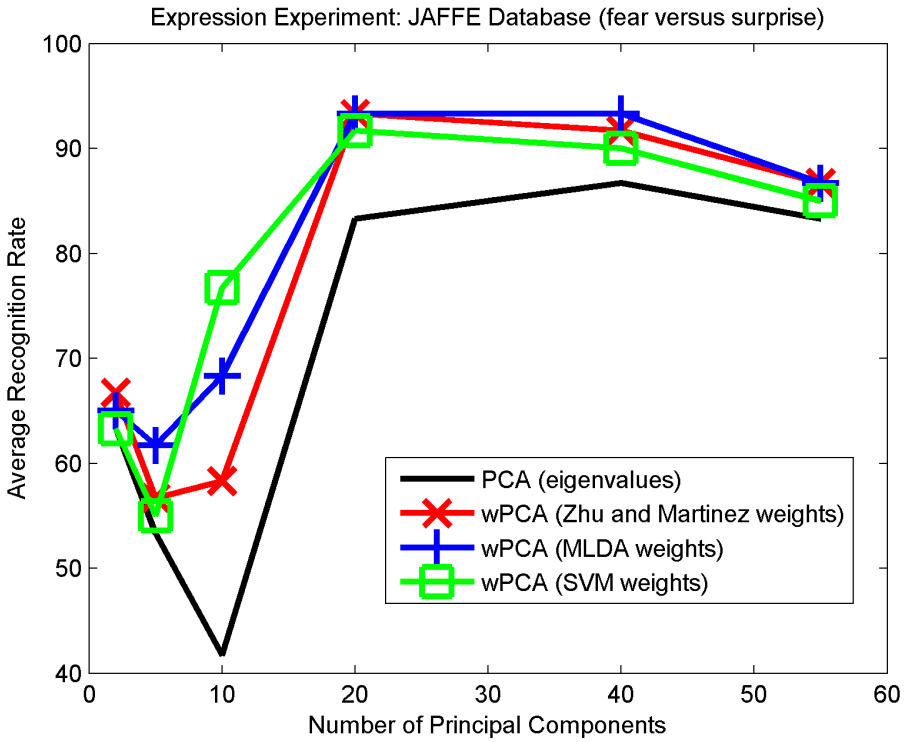

Figure 7: Fear versus surprise facial expression recognition performance of the spatially weighted PCA (wPCA) compared to the standard PCA using the JAFFE database. All the principal components retained have been selected by their corresponding largest eigenvalues.

making better predictions. Since the principal components in general, which correspond to the largest eigenvalues, minimize the mean square reconstruction over all choice of orthonormal basis vectors [8], we should expect a similar data compression of the weighted principal components compared to the standard ones, despite the additional normalization of the original data by inserting distinct spatial weights on individual pixels within the face images.

Figure 8 illustrates the first 7 principal components (from left to right), that is, the first 7 eigenvectors with the largest eigenvalues, of the FEI database using the following methods (from top to bottom): standard PCA (line 1), weighted PCA normalized by Zhu and Martinez gender (line 2) and expression (line 3) hyperplanes, weighted PCA normalized by MLDA gender (line 4) and expression (line 5) hyperplanes, and weighted PCA normalized by SVM gender (line 6) and expression (line 7) hyperplanes. It is possible to observe that using the weighted principal components the eigenvectors with the largest eigenvalues incorporate the specific domain knowledge into the data representation, highlighting, unlike the standard principal components (line 1 of Figure 8), the most important parts of the face images depending on the differences that we would like to describe between male and female samples (gender weighting hyperplanes, lines 2, 4 and 6 of Figure 8) or smiling and non-smiling samples (expression weighting hyperplanes, lines 3, 5 and 7 of Figure 8). Moreover, as shown in Figure 9, the weighted principal components have the same characteristic exponential decay of the standard principal components, explaining similar amount of the total variance for each principal component retained in both gender and expression experiments using the FEI database.

In fact, we should expect a reconstruction process based on the weighted principal components similar to the standard ones, but more efficient for making predictions specially on the first axes projection because of the spatial weights control on the individual pixels within the face images, as illustrated in Figures 10 and 11. These figures show a male smiling (top left image on Figure 10) and a female non-smiling (top left image on Figure 11) samples of the FEI database projected on the principal components eigenspace and reconstructed using 5, 10, 20, 40, 80, 160, 320 and all principal components (from left to right) by the following methods (from top to bottom): standard PCA (line 1), weighted PCA by Zhu and Martinez gender (line 2) and expression (line 3) hyperplanes, weighted PCA by MLDA gender (line 4) and expression (line 5) hyperplanes, and weighted PCA by SVM gender (line 6) and expression (line 7) hyperplanes.

It is possible to see that the eigensubspace composed of the weighted principal components, for all the hyperplanes considered, tends to reconstruct firstly, using as few components as possible, the most discriminant parts of the face images for predicting either gender or expression differences. For example, in Figure 10, since the image to be reconstructed is of a smiling sample, if we would like to predict solely the facial expression of this sample, no matter the identity of the subject or its gender, we would need only 5 principal components weighted by the Zhu and Martinez, MLDA or SVM expression hyperplanes (lines 3, 5 and 7 on Figure 10) to incorporate such expression discriminant information in the reconstruction process, unlike the standard PCA (line 1) that would need at least 20 components. An analogous behavior can be observed in Figure 11, but now considering a non-smiling sample and mainly the Zhu and Martinez and MLDA facial expression hyperplanes (lines 3 and 5 on Figure 11). Figure 12 summarizes these visual results by plotting the squared reconstruction error of the n-dimensional illustrated samples calculated as follows:

$$e_r(\mathbf{x}_i, \mathbf{x}_r) = \sum_{j=1}^{n} w_j (x_{ij} - x_{rj})^2, \tag{17}$$

Figure 8: The first 7 principal components (from left to right) of the FEI database using the following methods (from top to bottom): standard PCA (line 1), weighted PCA by Zhu and Martinez gender (line 2) and expression (line 3) hyperplanes, weighted PCA by MLDA gender (line 4) and expression (line 5) hyperplanes, and weighted PCA by SVM gender (line 6) and expression (line 7) hyperplanes.

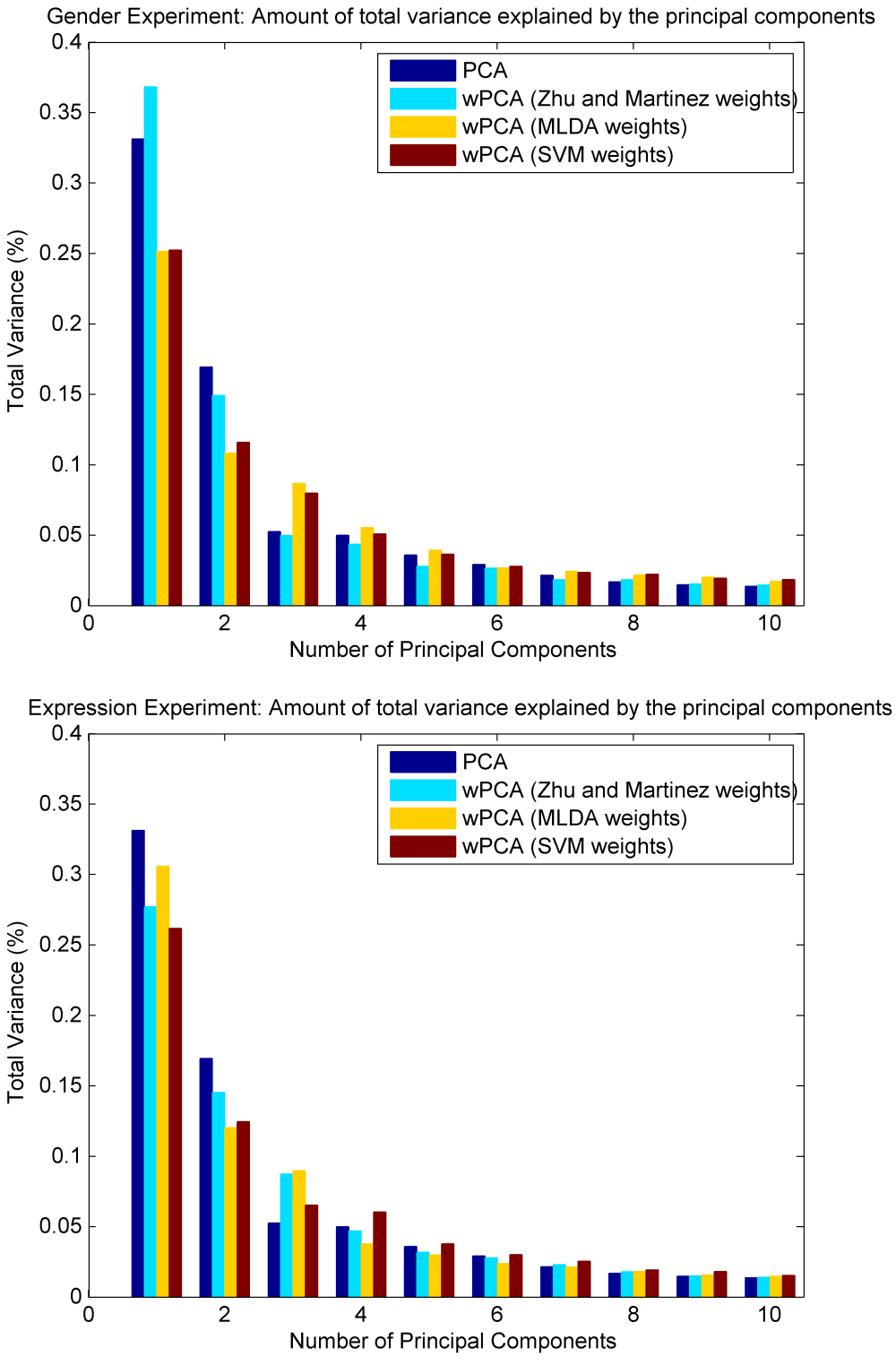

Figure 9: Amount of total variance explained by the standard and weighted principal components of the FEI database on the gender (top) and expression (bottom) experiments. It is possible to see that the weighted principal components have the same characteristic exponential decay of the standard principal components (only 10 principal components shown).

Figure 10: Reconstruction process of a male smiling sample (top left) using 5, 10, 20, 40, 80, 160, 320 and all principal components (from left to right) of the FEI database using the following methods (from top to bottom): standard PCA (line 1), weighted PCA by Zhu and Martinez gender (line 2) and expression (line 3) hyperplanes, weighted PCA by MLDA gender (line 4) and expression (line 5) hyperplanes, and weighted PCA by SVM gender (line 6) and expression (line 7) hyperplanes.

Figure 11: Reconstruction process of a female non-smiling sample (top left) using 5, 10, 20, 40, 80, 160, 320 and all principal components (from left to right) of the FEI database using the following methods (from top to bottom): standard PCA (line 1), weighted PCA by Zhu and Martinez gender (line 2) and expression (line 3) hyperplanes, weighted PCA by MLDA gender (line 4) and expression (line 5) hyperplanes, and weighted PCA by SVM gender (line 6) and expression (line 7) hyperplanes.

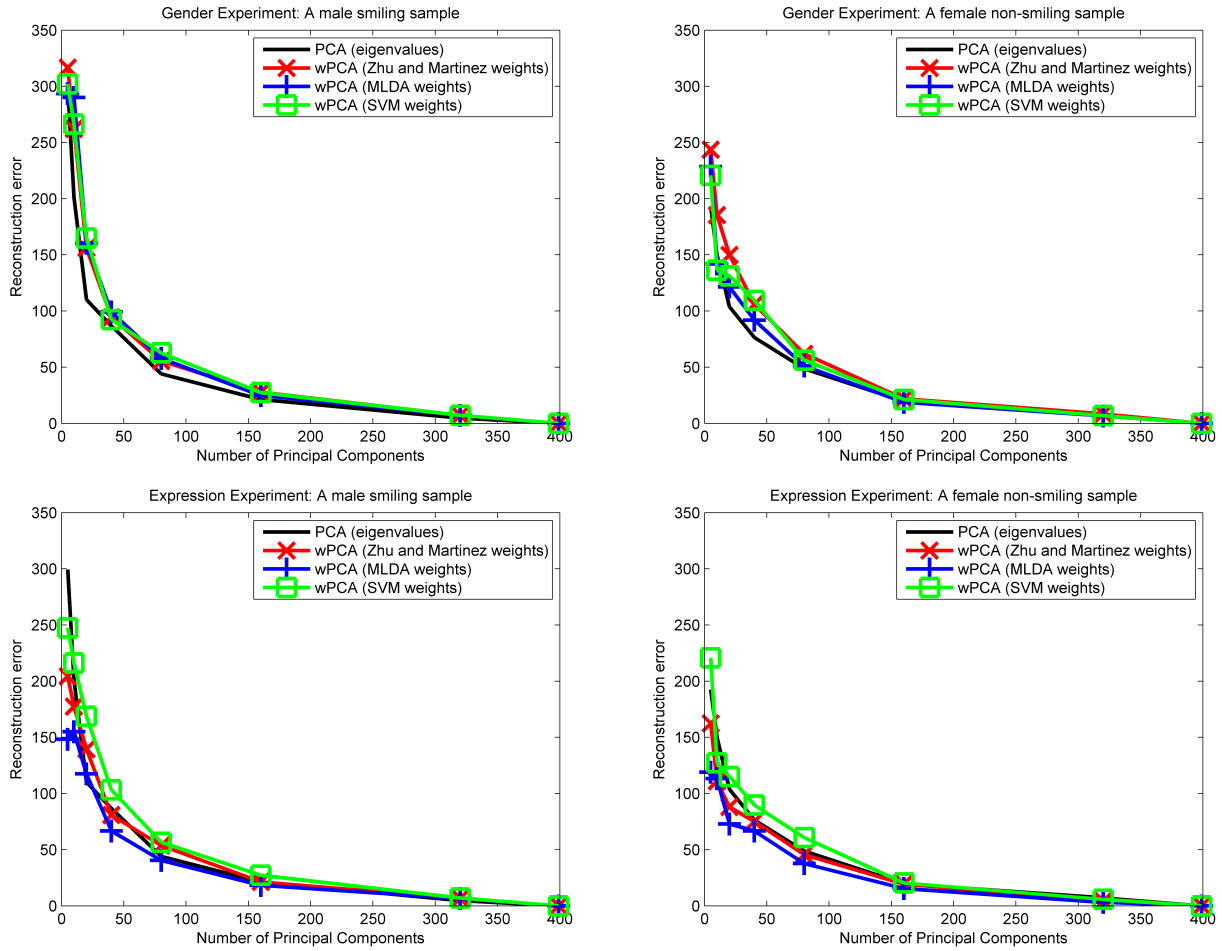

Figure 12: Squared reconstruction error e_r of the male smiling and female non-smiling samples using several principal components and the standard and weighted PCA projections.

where w_j are the spatial weights corresponding to the separating hyperplanes on the gender and expression discriminant tasks, $\mathbf{x}_i$ the male smiling or female non-smiling samples and $\mathbf{x}_r$ the respective reconstructed n-dimensional image vector using 5, 10, 20, 40, 80, 160, 320 and all principal components. For comparison, the reconstruction error of each pixel for the standard PCA projection has been weighted equally, that is, $w_j = \frac{1}{n}$ for the standard PCA.

All these results indicate that the weighted principal components are not only more efficient for making predictions but also achieve a reconstruction of the data as good as the standard ones taking into account explicitly the specific domain knowledge about the pattern recognition problem under investigation.

5 Discussion

Recently, we proposed a new ranking method for the principal components, called discriminant principal components [**?**], which is also based on the discriminant weights given by separating hyperplanes to determine among the standard principal components the most discriminant ones. Such approach does not deal with the problem of calculating new principal components but rather with the idea of selecting among the standard principal components the most discrimi-

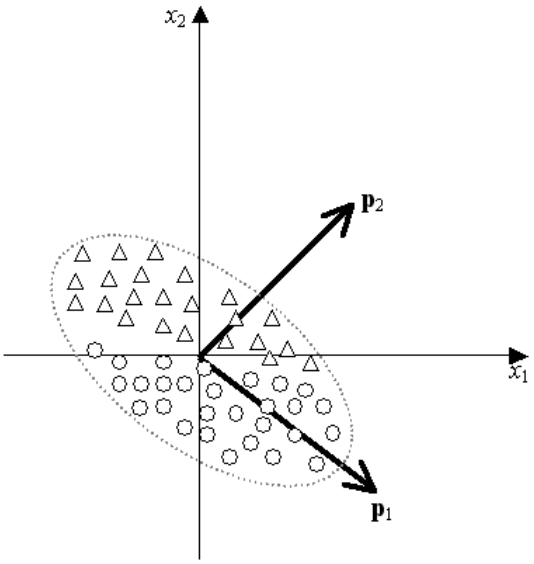

Figure 13: An hypothetical example that shows samples from two classes (depicted in two-dimensional points represented by triangles and circles) and the standard principal components $[\mathbf{p}_1, \mathbf{p}_2]$.

nant ones determined by how well they align with the specific separating hyperplane used [**?**]. In Figures 13-16, we describe geometrically the main difference between the spatially weighted and discriminant principal components.

Figures 13-16 illustrate an hypothetical example with samples from two classes (depicted in triangles and circles) along with the optimal corresponding separating hyperplane (dashed line in red). In Figure 13, we can see that the standard principal components $[\mathbf{p}_1, \mathbf{p}_2]$ are obtained, as well-known, by rotating the original coordinate axes until they coincide with the axes of the constant density ellipse described by all the samples, without taking into account any discriminant information available to separate the triangles from circles. The criterion of selecting the discriminant principal components [**?**] is based, however, on re-ranking the standard principal components that most correlate to $\mathbf{w} = [w_1, w_2]$, so the first discriminant principal component $\mathbf{p}_1^+$ should be equal to the second standard principal component $\mathbf{p}_2$, that is, $\mathbf{p}_1^+ = \mathbf{p}_2$ and, consequently, $\mathbf{p}_2^+ = \mathbf{p}_1$ because $|\mathbf{w} \cdot \mathbf{p}_2| > |\mathbf{w} \cdot \mathbf{p}_1|$, as illustrated in Figure 14. As described in the previous sections, the spatially weighted approach uses the discriminant weight of each original variable for finding a new orthonormal basis that are not necessarily composed of the standard principal components. That is, in this hypothetical example the influence of the variable deviations on the x_2 axis will be relatively magnified in comparison with x_1, because $|w_2| > |w_1|$ and $|w_1| + |w_2| = 1$. This is geometrically represented in Figure 15, where we indicate that the constant density ellipse will be expanded in the x_2 axis and shrunk in the x_1 axis. Therefore, the first spatially weighted principal component $\mathbf{p}_1^*$ would be significantly different from the standard and discriminant principal components, and much closer to the x_2 direction than x_1, as shown in Figure 16.

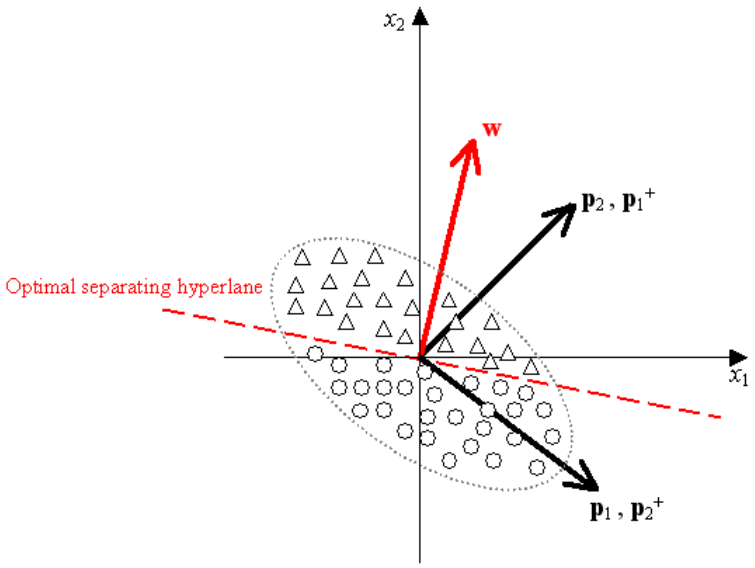

Figure 14: An hypothetical example that shows samples from two classes (depicted in two-dimensional points represented by triangles and circles) and the main geometric difference between the standard principal components $[\mathbf{p}_1, \mathbf{p}_2]$ and the discriminant ones $[\mathbf{p}_1^+, \mathbf{p}_2^+]$ that use the information alignment provided by $\mathbf{w}$ to re-rank the standard principal components.

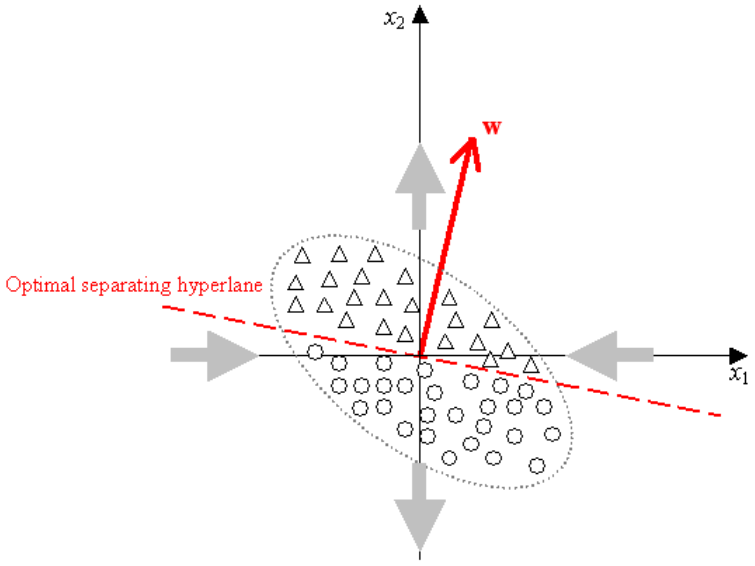

Figure 15: An hypothetical example that shows samples from two classes (depicted in two-dimensional points represented by triangles and circles) and the main geometric idea of the spatially weighting process that magnifies or shrinks the deviation of each variable depending on $\mathbf{w}$.

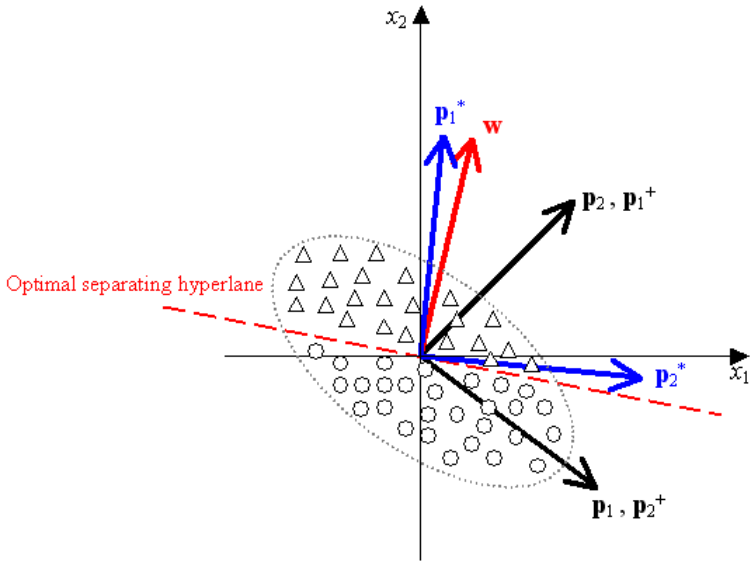

Figure 16: An hypothetical example that shows samples from two classes (depicted in two-dimensional points represented by triangles and circles) and the main geometric difference of standard $[\mathbf{p}_1, \mathbf{p}_2]$, discriminant $[\mathbf{p}_1^+, \mathbf{p}_2^+]$ and spatially weighted $[\mathbf{p}_1^*, \mathbf{p}_2^*]$ principal components.

6 Conclusion

In this paper, we proposed a supervised weighted PCA that incorporates domain knowledge and generates an embedding space that preserves the properties of dimensionality reduction and interpretability of the standard PCA, without jeopardizing its inherent straightforward and simple calculation. To evaluate the effectiveness of the weighted principal components proposed, we carried out several experiments using 2D frontal face images and different data sets. Our experimental results showed that the weighted principal components are not only more efficient for making predictions but also achieve a reconstruction of the data as good as the standard ones taking into account explicitly some labeled information available on the pattern recognition problem under investigation. We believe that such approach is a step forward towards the main objective of operating on high-dimensional and sparse input spaces where the problems of feature selection for reconstruction and learning, seen sometimes as mutually exclusive goals, are strongly connected.

References

[1] D. Cai, X. He, and J. Han. Semi-supervised discriminant analysis. In *Proceedings of the International Conference on Computer Vision (ICCV'2007)*, 2007.

[2] R. D. Cook and X. Yin. Dimension reduction and visualization in discriminant analysis (with discussion). *Australian and New Zealand Journal of Statistics*, 43:147–199, 2001.

[3] J. Pinto da Costa, I. Silva, and M. Eduarda Silva. Time dependent principal component analysis of time series data. In *IASC*, 2007.

[4] C. Davatzikos. Why voxel-based morphometric analysis should be used with great caution when characterizing group differences. *NeuroImage*, 23:17–20, 2004.

[5] P.A. Devijver and J. Kittler. *Pattern Classification: A Statistical Approach*. Prentice-Hall, 1982.

[6] D.L. Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. Technical report, Lecture delivered at the "Mathematical Challenges of the 21st Century" con- ference of The American Math. Society, Los Angeles, 2000.

[7] K. Forbes and E. Fiume. An efficient search algorithm for motion data using weighted pca. In *Proc. of the 2005 ACM SIGGRAPH/Eurographics Symp. on Comp. Anim.*, pages 67–76, 2005.

[8] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, New York, 1990.

[9] G. A. Giraldi, P. S. Rodrigues, E. C. Kitani, J. R. Sato, and C. E. Thomaz. Statistical learning approaches for discriminant features selection. *Journal of the Brazilian Computer Society*, 14(2):7–22, 2008.

[10] Isabelle Guyon and Andre Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, (3):1157–1182, 2003.

[11] J. Han and M. Kamber. *Data mining : concepts and techniques*. Morgan Kaufmann Publishers, 2001.

[12] T. Hastie, R. Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

[13] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313:504–507, 2006.

[14] R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis*. New Jersey: Prentice Hall, 1998.

[15] I.T. Jolliffe. Principal component analysis. *Springer Series in Statistics*, 2002.

[16] J.A. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction: Information Science and Statistics*. Springer, 2007.

[17] G. Lim and C. H. Park. Semi-supervised dimension reduction using graph-based discriminant analysis. *Inter. Conf. on Comp. and Inf. Tech.*, 1:9–13, 2009.

[18] M. J. Lyons, J. Budynek, and S. Akamatsu. Automatic classification of single facial images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(12):1357–1362, 1999.

[19] A. M. Martinez and M. Zhu. Where are linear feature extraction methods applicable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1934–1944, 2005.

[20] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901.

[21] P. J. Philips, H. Wechsler, J. Huang, and P. Rauss. The feret database and evaluation procedure for face recognition algorithms. *Image and Vision Computing*, 16(5):295–306, 1998.

[22] L. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces. *Journal of Optical Society of America*, 4(3):519–524, 1987.

[23] D. Skocaj, A. Leonardis, and H. Bischof. Weighted and robust learning of subspace representations. *Pattern Recogn.*, 40(5):1556–1569, 2007.

[24] M. Sugiyama. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *J. Mach. Learn. Res.*, 8:1027–1061, 2007.

[25] M. Sugiyama, T. Ide, S. Nakajima, and J. Sese. Semi-supervised local fisher discriminant analysis for dimensionality reduction. *Lecture Notes in Computer Science*, 5012:333–344, 2008.

[26] D. Swets and J. Weng. Using discriminants eigenfeatures for image retrieval. *IEEE Trans. Patterns Anal. Mach Intell.*, 18(8):831–836, 1996.

[27] C. E. Thomaz, E. C. Kitani, and D. F. Gillies. A maximum uncertainty lda-based approach for limited sample size problems - with application to face recognition. *Journal of the Brazilian Computer Society*, 12(2):7–18, 2006.

[28] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3:71–86, 1991.

[29] Vladimir N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, INC., 1998.

[30] L. Wang and Xiuju Fu. *Data mining with computational intelligence*. Springer, 2005.

[31] D. Zhang, Zhi-Hua Zhou, and S. Chen. Semi-supervised dimensionality reduction. In *Proc. of the 2007 SIAM Intern. Conf. on Data Mining*, 2007.

[32] Y. Zhang, Z. Qiu, and D. Sun. Palmprint identification using weighted pca feature. In *Proc. of the Inter. Conf. on Signal Processing*, pages 2112–2115, 2008.

[33] M. Zhu and T. J. Hastie. Feature extraction for nonparametric discriminant analysis. *Journal of Computational and Graphical Statistics*, 12:101–120, 2003.

[34] M. Zhu and A. M. Martinez. Subclass discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(8):1274–1286, 2006.

[35] M. Zhu and A. M. Martinez. Pruning noisy bases in discriminant analysis. *IEEE Transactions on Neural Networks*, 19(1):148–157, 2008.

[36] M. Zhu and Aleix M. Martinez. Selecting principal components in a two-stage lda algorithm. In *CVPR'06*, pages 132–137, June 2006.