Statistical Learning Approaches for Discriminant Features Selection

Gilson A. Giraldi[1], **Paulo S. Rodrigues**[4], **Edson C. Kitani**[2], **João R. Sato**[3] **and Carlos E. Thomaz**[5]

[1]Department of Computer Science
National Laboratory for Scientific Computing, LNCC
Petrópolis, Rio de Janeiro, Brazil
gilson@lncc.br

[2]Department of Electrical Engineering
University of São Paulo, USP
São Paulo, São Paulo, Brazil
ekitani@lsi.usp.br

[3]Institute of Radiology, Hospital das Clínicas (NIF-LIM44)
University of São Paulo, USP
São Paulo, São Paulo, Brazil
jrsatobr@gmail.com

[4]Department of Computer Science, [5]Department of Electrical Engineering
Centro Universitário da FEI, FEI
São Bernardo do Campo, São Paulo, Brazil
{psergio,cet}@fei.edu.br

Received 19 October 2007; accepted 19 May 2008

Abstract

Supervised statistical learning covers important models like Support Vector Machines (SVM) and Linear Discriminant Analysis (LDA). In this paper we describe the idea of using the discriminant weights given by SVM and LDA separating hyperplanes to select the most discriminant features to separate sample groups. Our method, called here as Discriminant Feature Analysis (DFA), is not restricted to any particular probability density function and the number of meaningful discriminant features is not limited to the number of groups. To evaluate the discriminant features selected, two case studies have been investigated using face images and breast lesion data sets. In both case studies, our experimental results show that the DFA approach provides an intuitive interpretation of the differences between the groups, highlighting and reconstructing the most important statistical changes between the sample groups analyzed.

Keywords: Supervised statistical learning, Discriminant features selection, Separating hyperplanes.

1. Introduction

Statistical learning theory explores ways of estimating functional dependency from a given collection of data. It covers important topics in classical statistics such as discriminant analysis, regression methods, and the density estimation problem [15, 11, 18]. Statistical learning is a kind of statistical inference, also called inductive statistics. It encompasses a rigorous qualitative theory to set the necessary conditions for consistency and convergence of the learning process as well as principles and methods based on this theory for estimating functions, from a small collection of data [14, 30].

Statistical inference has more than 200 years, including names like Gauss and Laplace. However, the systematic analysis of this field started only in the late 1920s. By that time, an important question to be investigated was finding a reliable method of inference, that means, to solve the problem: *Given a collection of empirical data originating from some functional dependency, infer this dependency* [30]. Therefore, the analysis of methods of statistical inference began with the remarkable works of Fisher (parametric statistics) and the theoretical results of Glivenko and Cantelli (convergence of the empirical distribution to the actual one) and Kolmogorov (the asymptotically rate of that convergence).

In the recent years, statistical learning models like Support Vector Machines (SVM) and Linear Discriminant Analysis (LDA) have played an important role for characterizing differences between a reference group of patterns and the population under investigation [1, 4, 13, 24, 21, 22, 23, 12]. In general, the basic pipeline to follow in this subject is: (a) Dimensionality reduction; (b) Choose a learning method to compute a separating hypersurface, that is, to solve the classification problem; (c) Reconstruction problem, that means, to consider how good a low dimensional representation might look like.

For instance, in image analyses it is straightforward to consider each data point (image) as a point in a n-dimensional space, where n is the number of pixels of each image. Therefore, dimensionality reduction may be necessary in order to discard redundancy and simplify further computational operations. The most known technique in this subject is the Principal Components Analysis (PCA) [10] which criterium selects the principal components with the largest eigenvalues [3, 16]. However, since PCA explains the covariance structure of all the data its most expressive components [20], that is, the first principal components with the largest eigenvalues, do not necessarily represent important discriminant directions to separate sample groups.

Starting from this observation, we describe in this work the idea of using the discriminant weights given by separating hyperplanes to select the most discriminant features to separate sample groups. The method, here called as Discriminant Feature Analysis or simply DFA, 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. In addition, the number of meaningful discriminant principal components is not limited to the number of groups. Furthermore, it can be applied to any feature space without the need of a pre-processing stage for dimensionality reduction. This is the key point we explore in this paper. Specifically, we show that DFA is able not only to determine the most discriminant features but also to rank the original features in ascending order of importance for classification.

To follow the key ideas of our proposal, we shall first consider classification and reconstruction problems in the context of statistical learning. We review the theory behind the cited methods, their common points, and discuss why SVM is in general the best technique for classification but not necessarily the best for extracting discriminant information. This will be discussed using face images and the separating hyperplanes generated by SVM [30, 14, 28] and a regularized version of LDA called Maximum uncertainty Linear Discriminant Analysis (MLDA) [27]. To further evaluate DFA on a data set not composed of images, we investigate a breast lesion classification framework, proposed in [17], that uses ultrasound features and SVM only. Following the radiologists knowledge, the feature space has been composed of the following attributes: area, homogeneity, acoustic shadow, circularity and protuberance. The DFA results confirm the experimental observations presented in [17] which indicate that features such as area, homogeneity, and acoustic shadow are more important to discriminate malign from benign lesions than circularity and protuberance.

The remainder of this paper is divided as follows. In section 2, we review SVM and LDA statistical learning approaches. Next, in section 3, we consider the classification and reconstruction problems from the viewpoint of SVM and LDA methods. Then, in Section 4, we present the DFA technique proposed in this work. Next, the experimental results used to help the discussion of the paper are presented, with two case studies: face image analysis (section 5) and breast lesion classification (section 6). Finally, in section 7, we conclude the paper, summarizing its main contributions and describing possible future works.

2. Statistical Learning Models

In this section we introduce and discuss some aspects of statistical learning theory related to Support Vector Machines and Linear Discriminant Analysis. The goal is to set a common framework for comparison and analysis of these separating hyperplanes on high dimensional and limited sample size problems. The material to be presented follows the references [30, 2, 14].

2.1. Support Vector Machines (SVM)

SVM [30] is primarily a two-class classifier that maximizes the width of the margin between classes, that is, the empty area around the separating hyperplane defined by the distance to the nearest training samples. It can be extended to multi-class problems by solving essentially several two-class problems.

Given a training set that consists of N pairs of $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2) \ldots (\mathbf{x}_N, y_N)$, where $\mathbf{x}_i$ denote the n-dimensional training observations and $y_i \in \{-1, 1\}$ are the corresponding classification labels. The SVM method [30] seeks to find the hyperplane defined by

$$f(\mathbf{x}) = (\mathbf{x} \cdot \mathbf{w}) + b = 0, \qquad (1)$$

which separates positive and negative observations with the maximum margin. The vector $\mathbf{w}$ and the scalar b

(threshold value) determine the orientation and position of the separating hyperplane. 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,

$$\mathbf{w}_{svm} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i, \tag{2}$$

where α_i are non-negative Lagrange coefficients obtained by solving a quadratic optimization problem with linear inequality constraints [2, 30]. Those training observations $\mathbf{x}_i$ with non-zero α_i lie on the boundary of the margin and are called support vectors.

2.2. Linear Discriminant Analysis (LDA)

The primary purpose of the LDA is to separate samples of distinct groups by maximizing their between-class separability while minimizing their within-class variability.

Let the scatter matrices between-class $\mathbf{S}_b$ and within-class $\mathbf{S}_w$ be defined, respectively, as

$$\mathbf{S}_b = \sum_{i=1}^{g} N_i (\overline{\mathbf{x}}_i - \overline{\mathbf{x}})(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})^T \tag{3}$$

$$\mathbf{S}_w = \sum_{i=1}^{g} (N_i - 1)\mathbf{S}_i = \sum_{i=1}^{g}\sum_{j=1}^{N_i} (\mathbf{x}_{i,j} - \overline{\mathbf{x}}_i)(\mathbf{x}_{i,j} - \overline{\mathbf{x}}_i)^T, \tag{4}$$

where $\mathbf{x}_{i,j}$ is the n-dimensional pattern (or sample) 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. The vector $\overline{\mathbf{x}}_i$ and matrix $\mathbf{S}_i$ are respectively the unbiased sample and sample covariance matrix of class i [10]. The grand mean vector $\overline{\mathbf{x}}$ is given by

$$\overline{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{g} N_i \overline{\mathbf{x}}_i = \frac{1}{N}\sum_{i=1}^{g}\sum_{j=1}^{N_i} \mathbf{x}_{i,j}, \tag{5}$$

where N is, as described earlier, the total number of samples, that is, $N = N_1 + N_2 + \ldots + N_g$. It is important to note that the within-class scatter matrix $\mathbf{S}_w$ defined in equation (4) is essentially the standard pooled covariance matrix $\mathbf{S}_p$ multiplied by the scalar $(N - g)$, where $\mathbf{S}_p$ can be written as

$$\begin{aligned}\mathbf{S}_p &= \frac{1}{N-g}\sum_{i=1}^{g}(N_i - 1)\mathbf{S}_i \\ &= \frac{(N_1 - 1)\mathbf{S}_1 + (N_2 - 1)\mathbf{S}_2 + \ldots + (N_g - 1)\mathbf{S}_g}{N-g}.\end{aligned} \tag{6}$$

The main objective of LDA is to find a projection matrix $\mathbf{W}_{lda}$ that maximizes the ratio of the determinant of the between-class scatter matrix to the determinant of the within-class scatter matrix (Fisher's criterium), that is,

$$\mathbf{W}_{lda} = \arg\max_{W} \frac{|\mathbf{W}^T\mathbf{S}_b\mathbf{W}|}{|\mathbf{W}^T\mathbf{S}_w\mathbf{W}|}. \tag{7}$$

The Fisher's criterium described in equation (7) is maximized when the projection matrix $\mathbf{W}_{lda}$ is composed of the eigenvectors of $\mathbf{S}_w^{-1}\mathbf{S}_b$ with at most $(g-1)$ nonzero corresponding eigenvalues [10, 8]. In the case of a two-class problem, the LDA projection matrix is in fact the leading eigenvector $\mathbf{w}_{lda}$ of $\mathbf{S}_w^{-1}\mathbf{S}_b$, assuming that $\mathbf{S}_w$ is invertible.

However, in limited sample and high dimensional problems, such as in face images analysis, $\mathbf{S}_w$ is either singular or mathematically unstable and the standard LDA cannot be used to perform the separating task. To avoid both critical issues, we have calculated $\mathbf{w}_{lda}$ by using a maximum uncertainty LDA-based approach (MLDA) that considers the issue of stabilizing the $\mathbf{S}_w$ estimate with a multiple of the identity matrix [26, 25, 27].

The MLDA algorithm can be described as follows:

1. Find the $\mathbf{\Phi}$ eigenvectors and $\mathbf{\Lambda}$ eigenvalues of $\mathbf{S}_p$, where $\mathbf{S}_p = \frac{S_w}{N-g}$;

2. Calculate the $\mathbf{S}_p$ average eigenvalue $\overline{\lambda}$, that is,

$$\overline{\lambda} = \frac{1}{n}\sum_{j=1}^{n}\lambda_j = \frac{Tr(\mathbf{S}_p)}{n}; \tag{8}$$

3. Form a new matrix of eigenvalues based on the following largest dispersion values

$$\begin{aligned}&\mathbf{\Lambda}^* = \\ &diag[max(\lambda_1, \overline{\lambda}), max(\lambda_2, \overline{\lambda}), \ldots, max(\lambda_n, \overline{\lambda})];\end{aligned} \tag{9}$$

4. Form the modified within-class scatter matrix

$$\mathbf{S}_w^* = \mathbf{S}_p^*(N-g) = (\mathbf{\Phi}\mathbf{\Lambda}^*\mathbf{\Phi}^T)(N-g). \tag{10}$$

The MLDA method is constructed by replacing $\mathbf{S}_w$ with $\mathbf{S}_w^*$ in the Fisher's criterium formula described in equation 7. It is based on the idea that in limited sample size and high dimensional problems where the within-class scatter matrix is singular or poorly estimated, the Fisher's linear basis found by minimizing a more difficult but appropriate *inflated* within-class scatter matrix would also minimize a less reliable *shrivelled* within-class estimate.

3. Classification versus Reconstruction

In this section, we consider classification and reconstruction problems from the viewpoint of LDA and SVM methods. As described in the previous section, both linear discriminant methods seek to find a decision boundary that separates data into different classes as well as possible.

The LDA solution is a spectral matrix analysis of the data and it depends on all of the data, even points far away from the separating hyperplane [14]. This can be seen by using the following example. Let $\mathbf{u}$ be a n-dimensional random vector with a mixture of two normal distributions with means $\overline{\mathbf{x}}_1$ and $\overline{\mathbf{x}}_2$, mixing proportions of p and $(1-p)$, respectively, and a common covariance matrix Σ. Then, it can be shown that the covariance matrix $\mathbf{S}$ of all the samples can be calculated as follows [3]:

$$\mathbf{S} = p(1-p)\mathbf{d}\mathbf{d}^T + \mathbf{\Sigma}, \tag{11}$$

where $\mathbf{d} = \overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2$, $\mathbf{S}_b = p(1-p)\mathbf{d}\mathbf{d}^T$ and $\mathbf{S}_w = \mathbf{\Sigma}$. Therefore, the Fisher's criterium in expression (7) becomes:

$$\mathbf{w}_{lda} = \arg\max_{w} \frac{\left|\mathbf{w}^T\mathbf{d}\mathbf{d}^T\mathbf{w}\right|}{\left|\mathbf{w}^T\mathbf{\Sigma}\mathbf{w}\right|}. \tag{12}$$

Thus, it is clear that the LDA solution depends on the class distributions, that is, the sample group means (or class prototypes) and covariance matrix (or spread of the sample groups). Consequently, LDA is less robust to gross outliers [14]. This is the reason why LDA may misclassify data points nearby the boundary of the classes.

On the other hand, the description of the SVM solution, summarized by expression (2), does not depend on the class distributions, focusing on support vectors which are the observations that lie on the boundary of the margin. In other words, SVM discriminant direction focuses on the data that are most important for classification, but such data are not necessarily the most important ones for extracting discriminant information between the sample groups. Figures 1 and 2 picture these aspects. They show a hypothetical data set composed of two classes and the separating planes obtained respectively by the LDA and SVM methods.

Figure 1 shows the LDA hyperplane which normal direction is clearly biased by the class distributions whereas the Figure 2 pictures SVM solution which, according to the optimality criterium behind SVM, searches for the plane that separates the subsets with maximal marging. Figure 1 illustrates the fact that LDA solution is less sensitive to the subtleties of group differences found in the frontiers of the classes than SVM which gives a zoom into these subtleties. This is the reason why SVM is more

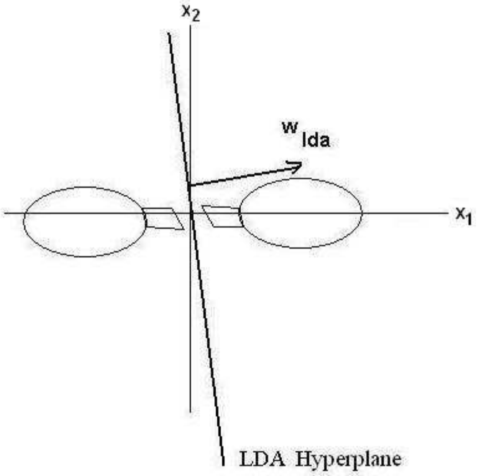


Figure 1. Hypothetical example: LDA separating hyperplane.

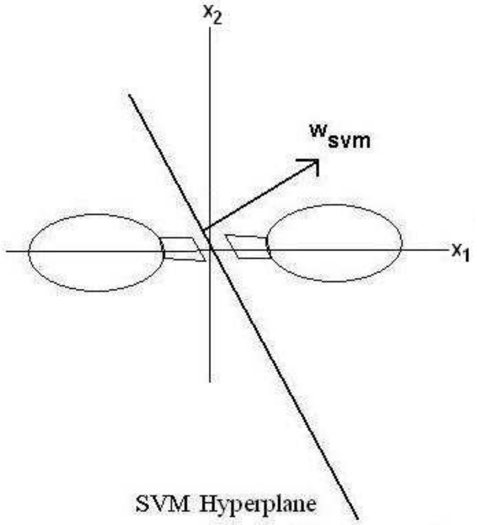


Figure 2. Hypothetical example: SVM separating hyperplane.

robust nearby the classification boundary of the classes achieving best recognition rates.

However, when considering reconstruction, things become different. Let a point $\mathbf{x}$ in the feature space be computed by the following expression:

$$\mathbf{x} = \overline{\mathbf{x}} + \delta \cdot \mathbf{w}, \quad \mathbf{w} \in \{\mathbf{w}_{lda}, \mathbf{w}_{svm}\}, \tag{13}$$

where $\delta \in \Re$ and $\overline{\mathbf{x}}$ is, for instance, the grand mean vector computed by expression (5). The LDA discriminant direction takes into account all the data because it maximizes the between-class separability while minimizing their within-class variability. This can be quantified by expression (12) which shows that LDA tries to collapse the classes into single points, their sample group means, as separated as possible. Therefore, when we set $\mathbf{w} = \mathbf{w}_{lda}$ in equation (13) we are moving along a direction that represents essentially the difference of the sample means normalized by the spread of the samples on the whole feature space when computing the $\mathbf{x}$ point. On the other hand, expression (2) computes the SVM discriminant direction $\mathbf{w}_{svm}$ through a linear combination of the support vectors. Therefore, SVM does not take into account the information about the class prototypes and spreads. Consequently, we expect a more informative reconstruction result when $\mathbf{w} = \mathbf{w}_{lda}$ than the one obtained by SVM discriminant direction, in terms of extracting group differences. Thus, Figures 1 and 2 are an attempt to represent these aspects in the sense that $\mathbf{w}_{lda}$ is closer than $\mathbf{w}_{svm}$ to the direction $\mathbf{d} = \overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2$.

4. Discriminant Features Analysis (DFA)

We approach the problem of selecting and reconstructing the most discriminant features as a problem of estimating a statistical linear classifier. Hence, an n-dimensional feature space $\{x_1, x_2, ..., x_n\}$ is defined and a training set, consisting of N measurements, is selected to construct both MLDA and SVM separating hyperplanes. However, we shall emphasize that any separating hyperplane can be used here.

Thus, assuming only two classes to separate, the initial training set is reduced to a data set consisting of N measurements on only 1 discriminant feature given by:

$$\begin{aligned} \tilde{y}_1 &= x_{11}w_1 + x_{12}w_2 + ... + x_{1n}w_n, \\ \tilde{y}_2 &= x_{21}w_1 + x_{22}w_2 + ... + x_{2n}w_n, \\ &\;\;\; ... \\ \tilde{y}_N &= x_{N1}w_1 + x_{N2}w_2 + ... + x_{Nn}w_n, \end{aligned} \tag{14}$$

where $\mathbf{w} = [w_1, w_2, ..., w_n]^T$ is the discriminant direction calculated by either MLDA or SVM approach, and $[x_{i1}, x_{i2}, ..., x_{in}]$, $i = 1, ..., N$ are the sample features.

We can determine the discriminant contribution of each feature by investigating the weights $[w_1, w_2, ..., w_n]^T$ of the corresponding discriminant direction $\mathbf{w}$. Weights that are estimated to be 0 or approximately 0 have negligible contribution on the discriminant scores $\tilde{y}_i$ described in equation (14), indicating that the corresponding features are not significant to separate the sample groups. In contrast, largest weights (in absolute values) indicate that the corresponding features contribute more to the discriminant score and consequently are important to characterize the differences between the groups.

Therefore, we select as the most important discriminant features the ones with the highest weights (in absolute values), that is, $|w_1| \geq |w_2| \geq ... \geq |w_n|$ described by either the MLDA separating hyperplane

$$\mathbf{w}_{mlda} = \arg\max_{w} \frac{|\mathbf{w}^T S_b \mathbf{w}|}{|\mathbf{w}^T S_w^* \mathbf{w}|} \tag{15}$$

or the SVM separating hyperplane, as described in equation (2) and repeated here as a reminder,

$$\mathbf{w}_{svm} = \sum_{i=1}^{N} \alpha_i y_i (\mathbf{x}_i). \tag{16}$$

In short, we are selecting among the original features the ones that are efficient for discriminating rather than representing the samples.

Once the statistical linear classifier has been constructed, we can move along its corresponding discriminant direction and extract the group differences captured by the classifier. Therefore, assuming that the spreads of the classes follow a Gaussian distribution and applying limits to the variance of each group, such as $\pm 3\sigma_i$, where σ_i is the standard deviation of each group $i \in \{1, 2\}$, we can move along $\mathbf{w}_{mlda}$ and $\mathbf{w}_{svm}$ and perform a discriminant features analysis of the data. Specifically, this mapping procedure is generated through expression (13), setting $\delta = j\sigma_i$, where $j \in \{-3, -2, -1, 0, 1, 2, 3\}$, $\overline{\mathbf{x}} = \overline{\mathbf{x}}_i$, and replacing $\mathbf{w}$ with $\mathbf{w}_{mlda}$ or $\mathbf{w}_{svm}$, that is:

$$\mathbf{x}_{i,j} = \overline{\mathbf{x}}_i + j\sigma_i \cdot \mathbf{w}, \qquad \mathbf{w} \in \{\mathbf{w}_{lda}, \mathbf{w}_{svm}\}. \tag{17}$$

This mapping procedure may work as a way of defining changes that come from "definitely group 1" and "definitely group 2" samples, and consequently investigating linear discriminant differences captured by the classifier that are beyond the average change described by each sample group.

5. Case Study 1: Face Images

We present in this section experimental results on face images analysis. These experiments illustrate firstly the reconstruction problem based on the most expressive principal components and then compare the discriminative information extracted by the MLDA and SVM linear approaches. Since the face recognition problem involves small training sets and a large number of features, common characteristics in several pattern recognition applications, and does not require a specific knowledge to interpret the differences between groups, it seems an attractive application to investigate and discuss the statistical learning methods studied in this work.

5.1. Face Database

We have used frontal images of a face database maintained by the Department of Electrical Engineering of FEI to carry out the experiments. The FEI face database contains a set of face images taken between June 2005 and March 2006 at the Artificial Intelligence Laboratory in São Bernardo do Campo, São Paulo, Brazil, with 14 images for each of 200 individuals - a total of 2800 images. All images are colorful and taken against a white homogenous background in an upright frontal position with profile rotation of up to about 180 degrees. Scale might vary about 10% and the original size of each image is 640x480 pixels. All faces are mainly represented by subjects between 19 and 40 years old with distinct appearance, hairstyle, and adorns. This database is publicly available for download on the following site http://www.fei.edu.br/∼cet/facedatabase.html.

To minimize image variations that are not necessarily related to differences between the faces, we first aligned all the frontal face images to a common template so that the pixel-wise features extracted from the images correspond roughly to the same location across all subjects. In this manual alignment, we have randomly chosen the frontal image of a subject as template and the directions of the eyes and nose as a location reference. For implementation convenience, all the frontal images were then cropped to the size of 360x260 pixels and converted to 8-bit grey scale. Since the number of subjects is equal to 200 and each subject has two frontal images (one with a neutral or non-smiling expression and the other with a smiling facial expression), there are 400 images to perform the experiments.

5.2. PCA Reconstruction Results

It is well-known that well-framed face images are highly redundant not only owing to the fact that the image intensities of adjacent pixels are often correlated but also because every individual has one mouth, one nose, two eyes, etc. As a consequence, we can apply dimensionality reduction in order to project an input image with n pixels onto a lower dimensional space without significant loss of information. Principal Components Analysis (PCA) is a feature extraction procedure concerned with explaining the covariance structure of a set of variables through a small number of linear combinations of these variables.

Thus, let an $N \times n$ training set matrix $\mathbf{X}$ be composed of N input face images with n pixels. This means that each column of matrix $\mathbf{X}$ represents the values of a particular pixel observed all over the N images. Let this data matrix $\mathbf{X}$ have covariance matrix $\mathbf{S}$ with respectively $\mathbf{P}$ and $\mathbf{\Lambda}$ eigenvector and eigenvalue matrices, that is,

$$\mathbf{P}^T \mathbf{S} \mathbf{P} = \mathbf{\Lambda}. \qquad (18)$$

It is a proven result that the set of m ($m \leq n$) eigenvectors of $\mathbf{S}$, which corresponds to the m largest eigenvalues, minimizes the mean square reconstruction error over all choices of m orthonormal basis vectors (Fukunaga, 1990). Such a set of eigenvectors that defines a new uncorrelated coordinate system for the training set matrix $\mathbf{X}$ is known as the principal components. In the context of face recognition, those $\mathbf{P}_{pca} = [\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_m]$ components are frequently called eigenfaces [29].

As the average face image is an n-dimensional point (n=360x260=93600) that retains all common features from the training sets, we could use this point to understand what happens statistically when we move along the principal components and reconstruct the respective coordinates on the image space. Analogously to the works by Cootes et al. [6, 5, 7], we have changed the average face image $\overline{\mathbf{x}}$ by reconstructing each principal component separately using the limits of $\pm\sqrt{\lambda_i}$, where λ_i are the corresponding largest eigenvalues. Specifically, we set $\delta = j\sqrt{\lambda_i}$, with $j \in \{-3, -2, -1, 0, 1, 2, 3\}$, in expression (13) and replace $\mathbf{w}$ with the principal directions, that is:

$$\mathbf{x}_{i,j} = \overline{\mathbf{x}} + j\sqrt{\lambda_i} \cdot \mathbf{p}_i, \quad i = 1, 2, \ldots, 7 \qquad (19)$$

where $\mathbf{p}_1, \mathbf{p}_2 \ldots, \mathbf{p}_7$ are the first seven most expressive principal components.

Figure 3 illustrates these transformations using the 400 frontal images available. As it can be seen, the first principal component (on the top) captures essentially the variations in the illumination and gender of the training samples. The second and third directions, in turn, model respectively variations related to grey-level of the faces and hair, and the shape of the head. Moreover, the fourth principal component captures the variation in the shape of the head as well as in the length of hair and in its grey-level. In contrast, the fifth most expressive component

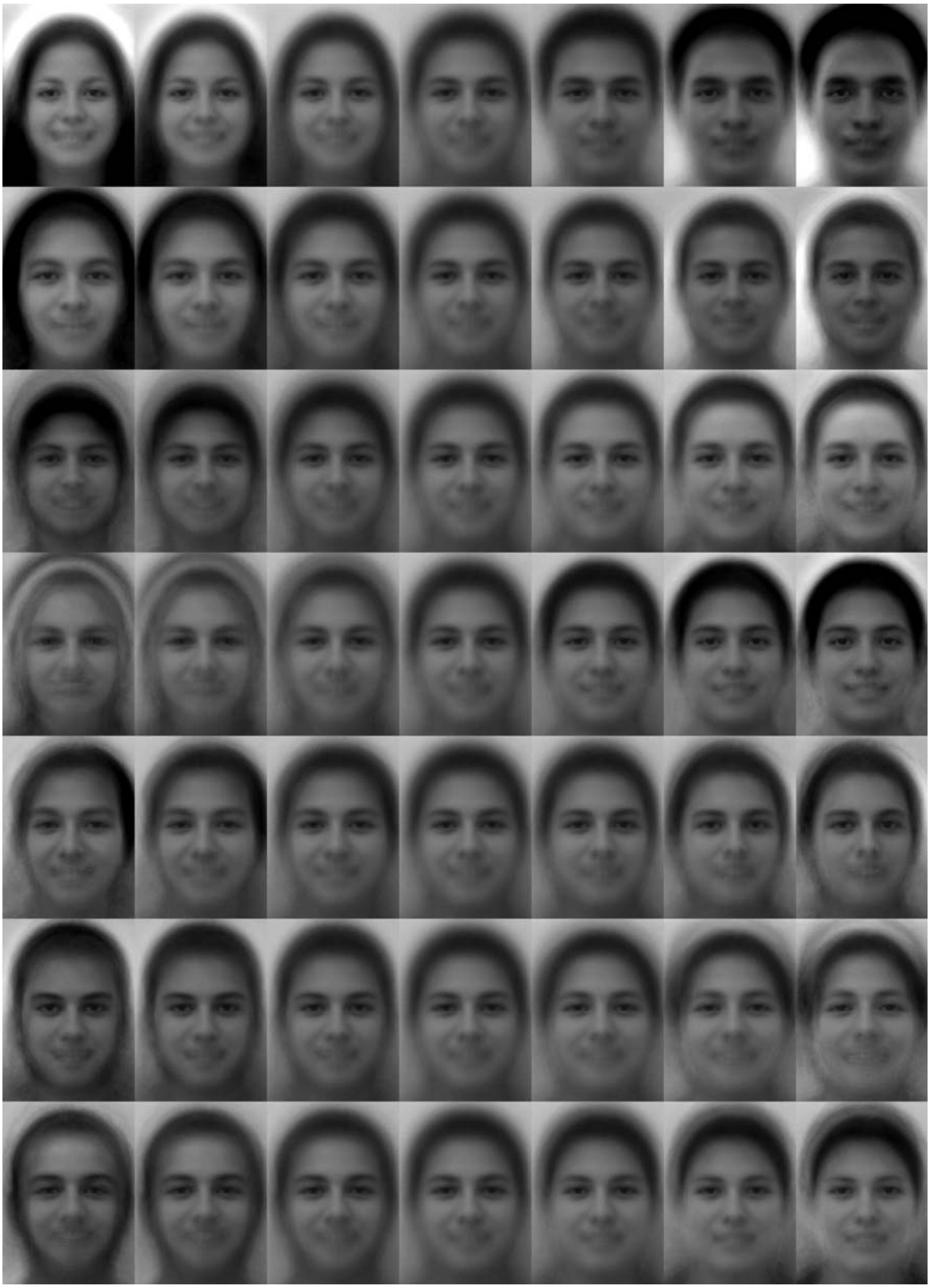


Figure 3. Reconstruction of the PCA most expressive components, i.e., from top to bottom, the first seven principal components with the largest eigenvalues in descending order. Each row i of images represents the following reconstruction defined by equation (19): $[\mathbf{x}_{i,-3}, \mathbf{x}_{i,-2}, \mathbf{x}_{i,-1}, \mathbf{x}_{i,0}, \mathbf{x}_{i,+1}, \mathbf{x}_{i,+2}, \mathbf{x}_{i,+3}]$, where $i = 1, 2, \ldots, 7$.

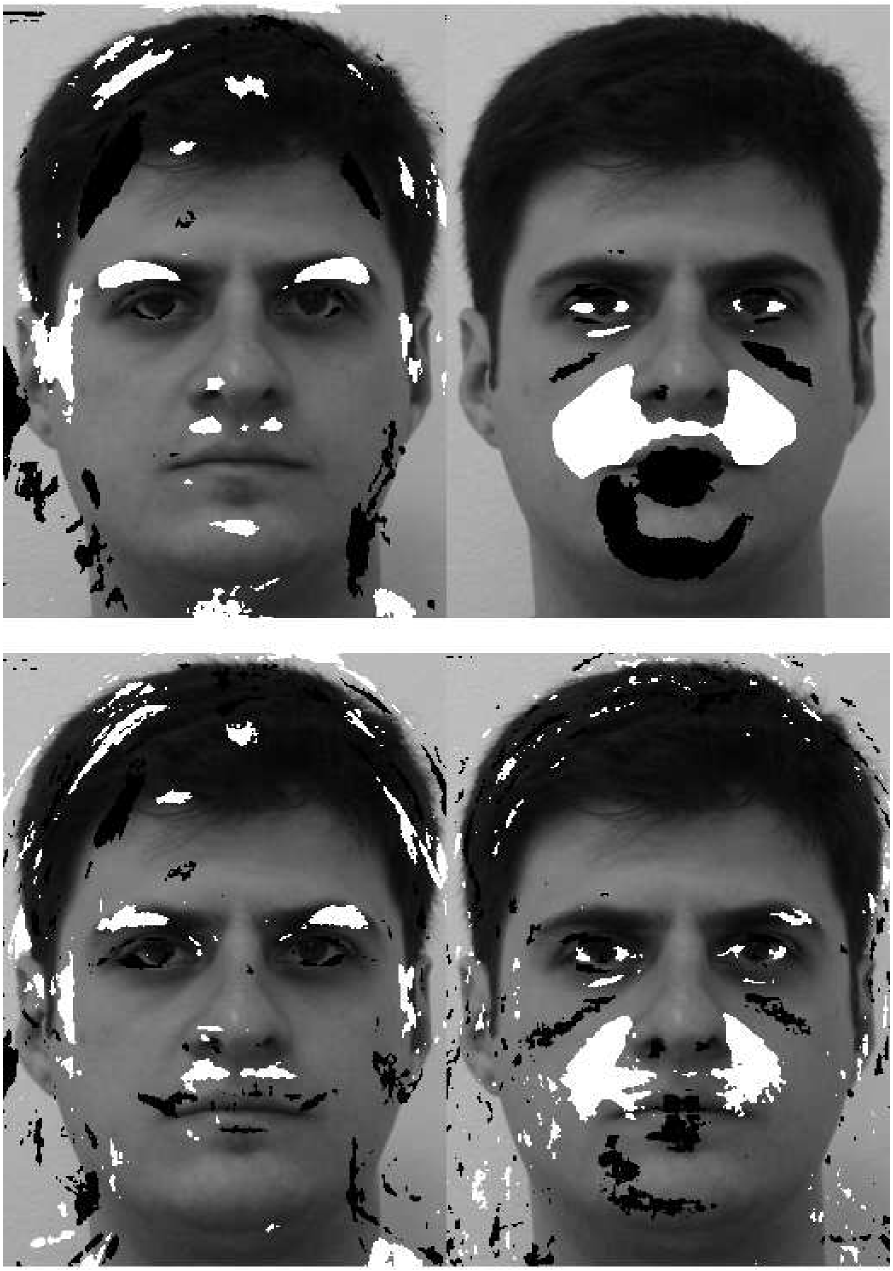


Figure 4. MLDA (top) and SVM (bottom) most discriminant pixels: (left) comparing the gender intensity changes; (right) comparing the expression intensity changes.

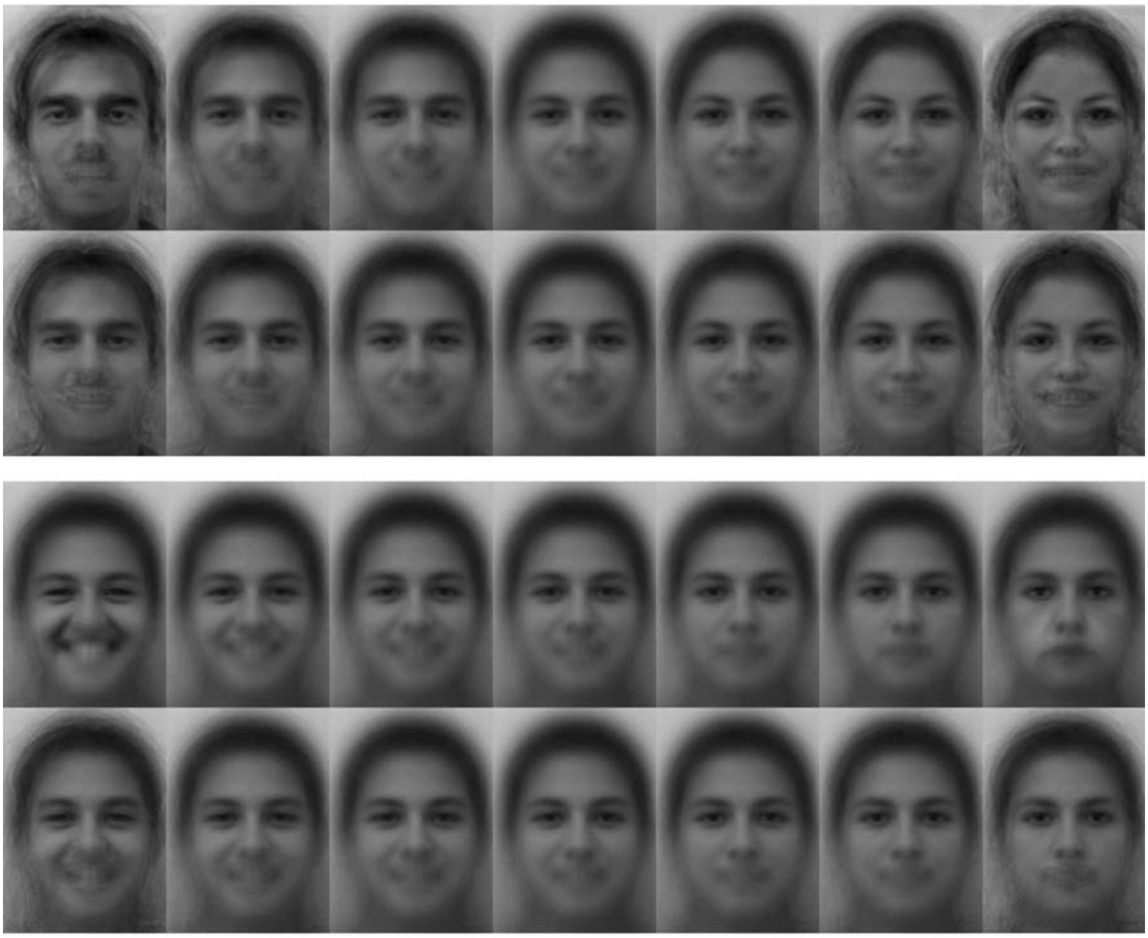


Figure 5. Reconstruction of the discriminant features captured by the MLDA and SVM statistical learning approaches. First two rows represent the gender experiments: MLDA (top) and SVM (bottom). Last two rows correspond to the facial expression experiments: MLDA (top) and SVM (bottom). From left (group 1 of male or smiling samples) to right (group 2 of females or non-smiling samples) each row of images represents the following reconstruction defined by equation (17): $[\mathbf{x}_{1,-3}, \mathbf{x}_{1,0}, \mathbf{x}_{1,+1}, \frac{\mathbf{x}_{1,0}+\mathbf{x}_{2,0}}{2}, \mathbf{x}_{2,-1}, \mathbf{x}_{2,0}, \mathbf{x}_{2,+3}]$.

describes some profile changes of the subjects that cannot be characterized as either gender or expression variation. The last two most expressive components capture other variations that are related to male and female differences such as the presence or absence of beard around the cheeks and chin regions.

As we should expect, these experimental results show that PCA captures features that have a considerable variation between all training samples, like changes in illumination, gender, and head shape. However, if we need to identify specific changes such as the variation in facial expression solely, PCA has not proved to be a useful solution for this problem. As it can be seen in Figure 3, although the fourth principal component models some facial expression variation, this specific variation has been subtly captured by other principal components as well including other image artifacts. Likewise, as Figure 3 illustrates, although the first principal component models gender variation, other changes have been modeled concurrently, such as the variation in illumination. In fact, when we consider a whole grey-level model without landmarks to perform the PCA analysis, there is no guarantee that a single principal component will capture a specific variation alone, no matter how discriminant that variation might be.

5.3. Information Extraction and Classification Results

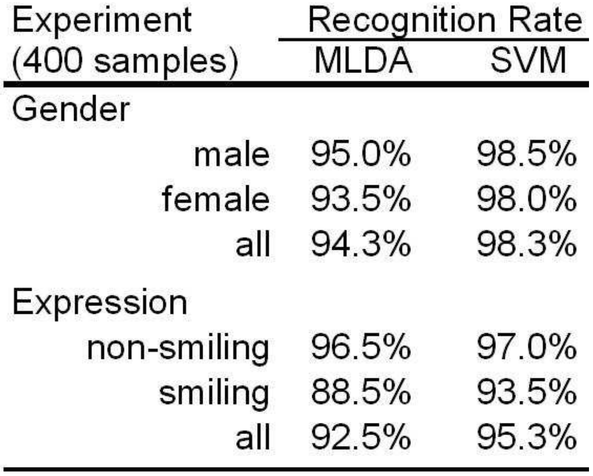
We have carried out the following two-group statistical analyzes: female versus male (gender) experiments, and non-smiling versus smiling (expression) experiments. We have composed the gender training set of 200 frontal female images, i.e. a mixture of non-smiling and smiling female images, and 200 analogous frontal male images. For the expression experiments, we have used the 200 frontal non-smiling images available, i.e. a mixture of female and male images, and their respective frontal smiling images. The idea of the first discriminant experiment is to evaluate the statistical learning approaches on a discriminant task where the differences between the groups are evident. The second experiment poses an alternative analysis where there are subtle differences between the groups.

Before evaluating the classification performance of the separating hyperplanes, we first analyze the linear discriminant features extracted by the MLDA and SVM statistical learning methods. Since the separating hyperplanes have been calculated on the PCA feature space, DFA can determine the MLDA and SVM discriminant contribution of each pixel on the original image space by multiplying $\mathbf{w}_{mlda}$ and $\mathbf{w}_{svm}$ by the transpose of the principal components matrix $\mathbf{P}_{pca}$.

Figure 4 shows the spatial distribution of the discriminant pixels extracted by each separating hyperplane superimposed on the template image used to align all the frontal faces. We highlight only the pixels which correspond to the 5% largest (in absolute values) positive and negative weights. We can see clearly that by exploring the separating hyperplane found by the statistical learning approaches and ranking their most discriminant pixels we are able to identify features that most differ between the group samples, such as: hair, eyebrow, eyes, nose, upper lip, chin and neck for the gender experiments; and eyes, shadow, cheek, upper lip and mouth for the facial expression experiments. As we should expect, changes in facial expression are subtler and more localized than the gender ones. Moreover, it is important to note that the discriminant features vary depending on the separating hyperplane used. We observe that the discriminant features extracted by MLDA are more informative and robust for characterizing group-differences than the SVM ones.

Figure 5 summarizes the reconstruction results captured by the multivariate statistical classifiers using all the gender and facial expression training samples. Specifically, these images are generated through expression (17), replacing $\mathbf{w}$ with $\mathbf{w}_{mlda}$ and $\mathbf{w}_{svm}$. As it can be seen, both MLDA and SVM hyperplanes similarly extracts the gender group differences, showing clearly the features that mainly distinct the female samples from the male ones, such as the size of the eyebrows, nose and mouth, without enhancing other image artifacts. Looking at the facial expression spatial mapping, however, we can visualize that the discriminative direction found by the MLDA has been more effective with respect to extracting group-differences information than the SVM one. For instance, the MLDA most discriminant direction has predicted facial expressions not necessarily present in our corresponding expression training set, such as the "definitely smiling" or may be "happiness" status and "definitely non-smiling" or may be "anger" status represented respectively by the left most and right most images in the third row of Figure 5.

Finally, Table 1 shows the leave-one-out recognition rates of the MLDA and SVM classifiers on the gender and facial expression experiments. As it can be seen, in both experiments SVM achieved the best recognition rates showing higher classification results than the MLDA approach. These results confirm the fact that SVM is a more robust technique for classification than MLDA, as already pointed out in section 3.



Experiment	Recognition Rate	
(400 samples)	MLDA	SVM
Gender		
male	95.0%	98.5%
female	93.5%	98.0%
all	94.3%	98.3%
Expression		
non-smiling	96.5%	97.0%
smiling	88.5%	93.5%
all	92.5%	95.3%

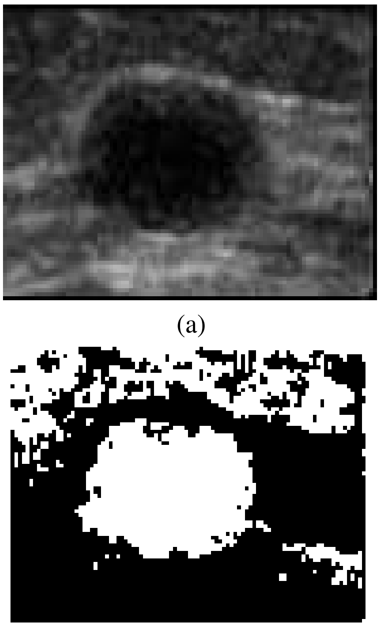
Table 1. Classification results of the MLDA and SVM separating hyperplanes.

6. Case Study 2: Breast Lesion

In [17], authors have proposed an automatic methodology for breast lesion classification in ultrasound images based on the following five-step framework: (a) Non-extensive entropy segmentation algorithm; (b) Morphological cleaning to improve segmentation result; (c) Accurate boundary extraction through level set framework; (d) Feature extraction; (e) SVM non-linear classification using the breast lesion features as inputs.

6.1. Five-Step Ultrasound Framework

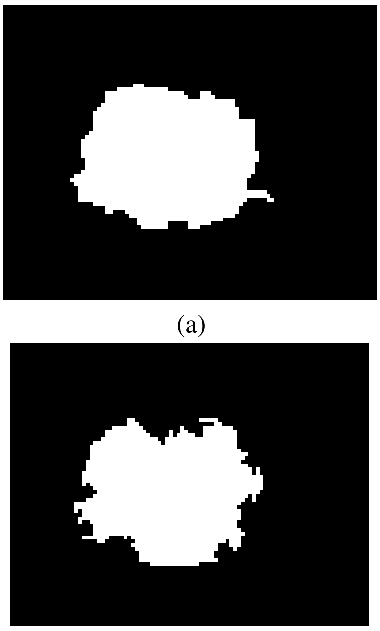
The first step of the framework for breast lesion analysis performs an initial segmentation of the ultrasound image using a generalization of the well known Boltzman-Gibbs-Shannon entropy. In [17] it was presented an algorithm, called NESRA (Non-Extensive Segmentation Recursive Algorithm) to detect the main regions of the ultrasound images (say, the tumor and the background ones) as well as the narrow region around the tumor. These regions are fundamental to further extracting the tumor features. Figure 6 shows an image example of an original benign lesion used in this work (on the top) and the corre-



(a)

(b)

Figure 6. (a) Original ultrasound benign image; (b) NESRA segmentation.



(a)

(b)

Figure 7. (a) ROI after morphological step; (b) final ROI after the level set approach.

sponding result after the non-extensive segmentation with the NESRA algorithm (on the bottom). The justification of why using a non-extensive segmentation algorithm for ultrasound images can be found in [17] and references therein.

As described in the framework, in the second step we have used a morphological chain approach in order to extract the region of interest (ROI) from the background. This was accomplished through the following rule. Considering the binary image generated by NESRA (e.g Figure 6-b), let α and β be the total ROI's area and the total image area, respectively. If $\alpha \geq \xi\beta$ an erosion is carried out and if $\alpha \leq \delta\beta$ a dilation is performed. Assuming that the ROI has a geometric point near to the image center, we apply a region growing algorithm which defines the final ROI's boundary. In [17] it was set $\xi = 0.75$ and $\delta = 0.25$ to extract most the ROIs. The result of this morphological rule applied in the image of Figure 6-b is illustrated in Figure 7-a. As it can be seen, the region generated by the morphological chain rule is a coarse representation of the lesion region. Then, we have applied a level set framework [17] using as initialization this region's boundary [19]. The result can be seen in Figure 7-b, which was accomplished with only 10 iterations of the level set approach.

The result of the lesion extraction illustrated in Figure 7 was used as input to calculate the tumor features commonly used by radiologists in diagnosis. Then, the next step is the feature extraction of the ROI. In the work presented in [17], three radiologists stated five features which have high probability to work well as a discriminator between malignant and benign lesions. Then, we have used these features and tested them in order to achieve the best combination in terms of performance. The feature space has been composed of the following attributes:

- Area (AR): The first feature considered is the lesion area. As indicated by the radiologists, since malignant lesions generally have large areas in relation to benign ones, this characteristic might be an important discriminant feature. We have normalized it by the total image area.
- Circularity (CT): The second characteristic is related to the region circularity. Since benign lesions generally have more circular areas compared with the malignant ones, this can also be a good discriminant feature. Then, we have taken the ROI's geometric center point and compute the distance from each boundary point (x_i, y_i) to it. We should expect that malignant lesions tend to have high standard deviations of the average distances in relation to the benign ones. Also, this feature is normalized by total image area.
- Protuberance (PT): The third feature is the size distribution of the lobes in a lesion. A boundary's lobe is a protuberant region on the boundary. We have computed the convex hull of the ROI and the lobe

as a protuberance between two valleys. The lobe areas are computed and only those greater than 10% of the lesion area are considered. This feature is taken as the average area of the lobes. According to the radiologists, we might expect that malignant lesions have higher average area than benign ones.

- Homogeneity (HO): The next feature is related to the homogeneity of the lesion. Malignant lesions tend to be less homogeneous than benign ones. Then, we take the Boltzman-Gibbs-Shannon entropy – taken over the gray scale histogram – relative to the maximum entropy as the fourth discriminant feature. In this case, we should expect that as higher the relative entropy less homogeneous is the lesion region and, consequently, higher is the chance to be a malign lesion.

- Acoustic Shadow (AS): The last feature is related with a characteristic called acoustic shadow. In benign lesions there are many water particles and, as a consequence, dark areas below such lesions are likely to be detected. On the other hand, when the lesion is more solid (a malignant characteristic), there is a tendency in forming white areas below it. We have computed the relative darkness between both areas (lesion's area and area below the lesion) and have taken it as the fifth lesion feature.

These features are the input to a SVM classifier that separates the breast lesions between malignant and benign types. The applied SVM utilizes B-spline as a kernel in its framework. In [17], authors justified the use of a B-Spline as a kernel for the SVM by comparing its performance with polynomial and exponential kernels. Additionally, in [17], ROC analyzes of several combinations of the five-feature set have been performed to determine the best recognition performance of the framework. Although the experimental results reported in [17] have shown that area, homogeneity, and acoustic shadow gives the best classification rates, no theoretical justification was presented in order to select a specific subset of the original feature space for optimum information extraction and classification performance.

6.2. Information Extraction and Classification Results

We repeat the same experiments carried out in [17], which have used a 50 pathology-proven cases database (20 benign and 30 malignant) to evaluate our DFA method on the five-step ultrasound framework previously described. Each case is a sequence of 5 images of the same lesion. Thus, we tested 100 images of benign lesion and 150 of malignant ones, that is, a total of 250 difference case.

Since the SVM separating hyperplane has been calculated on the original feature space, DFA can determine the discriminant contribution of each feature by investigating the weights of the most discriminant direction found by the SVM approach. Table 2 lists the features in decreasing order of discriminant power (in absolute values) selected by the SVM separating hyperplane using all the samples available. As it can be seen, SVM has selected the AR feature as the most discriminant feature, followed by AS, HO, CT and PT. In other words, we should expect a better performance of the classifier when using, for instance, two features only, if we select the pair of features (AR, AS) rather than (CT, PT).

1	Area (AR)
2	Acoustic Shadow (AS)
3	Homogeneity (HO)
4	Circularity (CT)
5	Protuberance (PT)

Table 2. SVM most discriminant features in decreasing order.

Analogously to the previous face experiments, the other main task that can be carried out by the DFA approach is to reconstruct the most discriminant feature described by the SVM separating hyperplane. Figure 8 presents the SVM most discriminant feature of the five-feature dataset using all the examples as training samples. It displays the differences on the original feature space captured by the classifier that change when we move from one side (malignant or group 1) of the dividing hyperplane to the other (benign or group 2). Specifically, these variations are generated through expression (17), replacing $\mathbf{w}$ with $\mathbf{w}_{svm}$. We can see clearly differences in the AR as well as AS and HO features. That is, the changes on the AR, AS, and HO features are relatively more significant to discriminate the sample groups than the CT and PT features. Additionally, Figure 8 illustrates that when we move from the definitely benign samples (on the right) to the definitely malign samples (on the left), we should expect an relative increase on the lesion area (AR), and a relative decrease on the acoustic shadow (AS) and homogeneity (HO) of the lesion. All these results are plausible and provide a quantitative measure to interpreting the discriminant importance and variation of each feature in the classification experiments.

Following the discriminant order and importance suggested in Table 2 and Figure 8 respectively, we can guide our classification experiments by combining the features according to those which improve the groups separation. Then, we carried out further classification experiments according to the following features combination:

1. AR + AS

2. AR + AS + HO

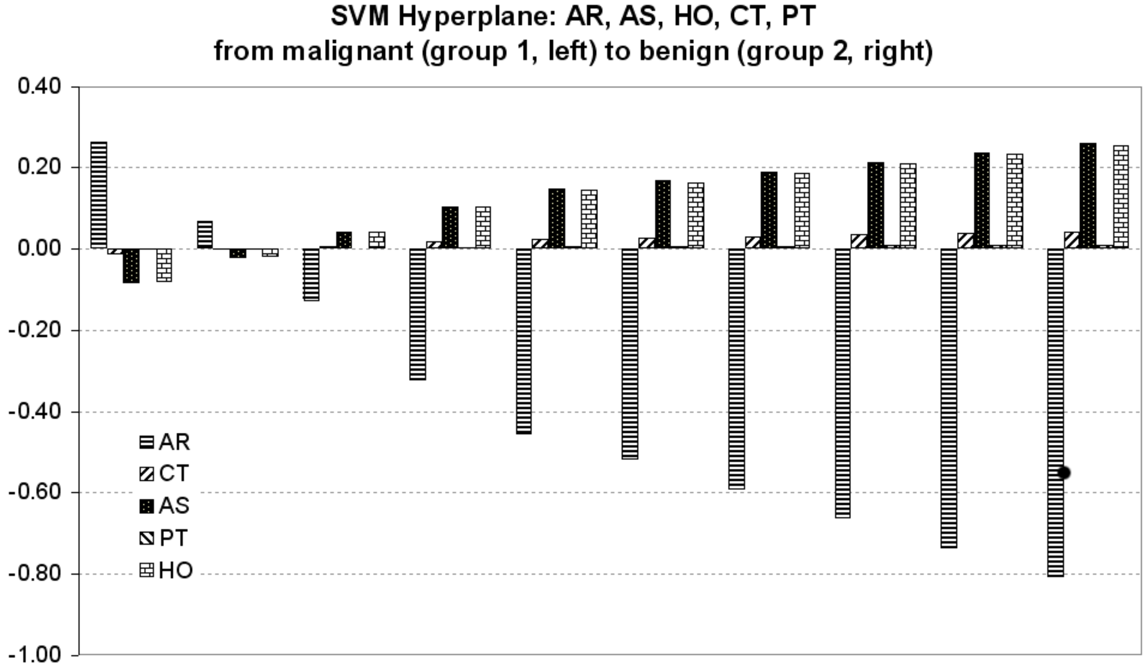


Figure 8. Reconstruction of the discriminant features captured by the SVM statistical learning approach on the five-step framework. From left (group 1 of malignant samples) to right (group 2 of benign samples) each five-set of clustered bars represents the following reconstruction defined by equation (17): $[\mathbf{x}_{1,-3}, \mathbf{x}_{1,-2}, \mathbf{x}_{1,-1}, \mathbf{x}_{1,0}, \frac{\mathbf{x}_{1,0}+\mathbf{x}_{2,0}}{2}, \mathbf{x}_{2,-1}, , \mathbf{x}_{2,0}, \mathbf{x}_{2,+1}, \mathbf{x}_{2,+2}, \mathbf{x}_{2,+3}]$.

3. AR + AS + HO + CT
4. AR + AS + HO + CT + PT

We have adopted the same cross-validation strategy carried out in [17] to evaluate these classification experiments. That is, the ultrasonic images are firstly divided randomly into five groups. We first set the first group as a test set and use the remaining four groups to train the SVM. After training, SVM is then tested on the first group. Then, we set the second group as a testing group and the remaining four groups as training set, and then SVM is tested on the second. This process is repeated until all the five groups have been set in turn as test sets.

In order to evaluate our results, we have used the Receiver Operating Characteristic (ROC) curve, which is a useful graph for organizing classifiers and visualizing their performance. Also, it is the most commonly used tool in medical decision tasks, and in recent years have been used increasingly in machine learning and data mining research. A nice recent review and discussion about the ROC analysis can be found in [9]. Briefly, ROC curves are two-dimensional graphs where true positive (TP) rate is visualized on the vertical axis and false positive (FP) rate is visualized on the horizontal axis. As such, a ROC curve tries to show relative tradeoffs between benefits (true positives) and costs (false positives) of a measuring. Other two indexes are true negative (TN) and false negative (FN). Under these four indexes, we can set several other indexes as follows:

- Accuracy = $(TP + TN)/(TP + TN + FP + FN)$
- Sensitivity = $TP/(TP + FN)$
- Specificity = $TN/(TN + FP)$
- Positive Predictive Value (PPV) = $TP/(TP + FP)$
- Negative Predictive Value(NPV) = $TN/(TN+FN)$

Together, all these five indexes guarantee a perfect classifier measuring. The commonly strategy is plotting the sensitivity as a function of the (1 - specificity) (see, for instance [9]) as a ROC curve. As larger the area under the ROC curve (that is, Az area), together with high accuracy, PPV and NPV, better is the classifier performance in terms of separating the sample groups.

The ROC curves of these experiments are shown in Figure 9. As it can be seen, the best combination which yields the largest Az value is clearly the one composed of the features AR, AS and HO, with $Az = 92\%$, as suggested previously by our discriminant features analysis. For this combination of features, our statistical framework achieved the following accuracy, PPV and NPV rates, respectively: 95%, 92% and 98%. The other combinations show lower Az values and worst sensitivity and (1 - specificity) ratios compared to the AR+AS+HO features set.

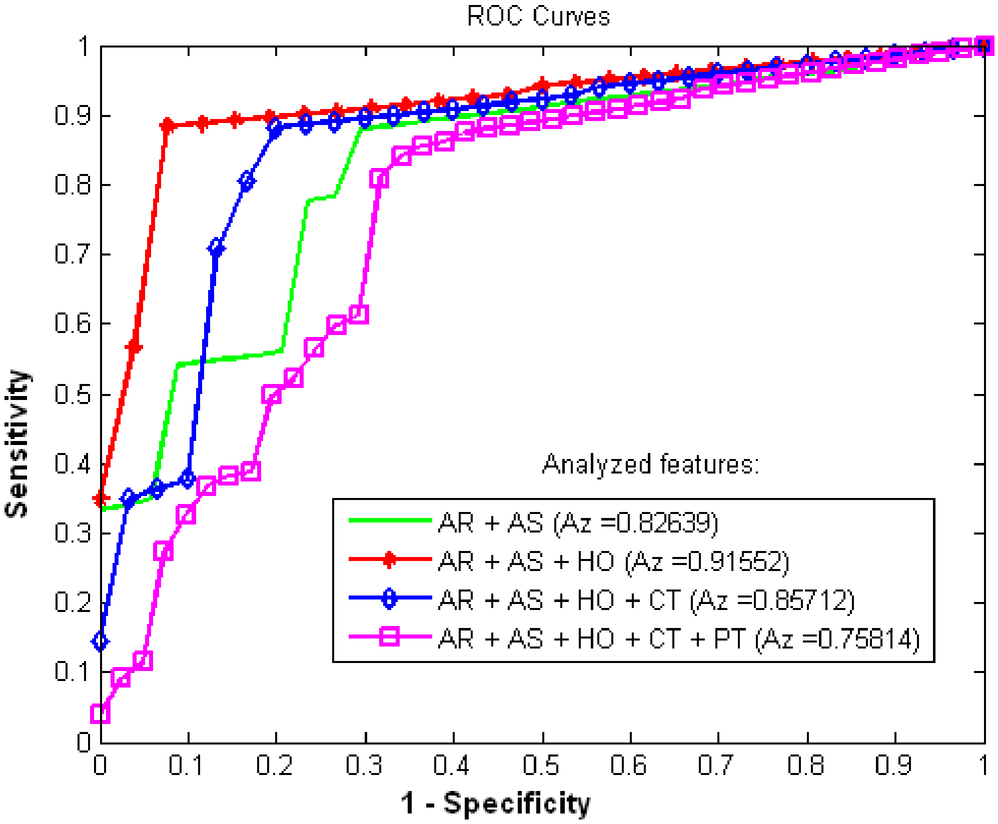


Figure 9. ROC curves for different combinations of the discriminant features in SVM classification of breast lesions.

7. Conclusion

In this paper, we described and implemented a method of discriminant features analysis based on sample group-differences extracted by separating hyperplanes. This method, called here simply as DFA, is based on the idea of using the discriminant weights given by statistical linear classifiers to select and reconstruct among the original features the most discriminant ones, that is, the original features that are efficient for discriminating rather than representing all samples. The two case studies carried out in this work using face images and breast lesion data sets indicated that discriminant information can be efficiently captured by a linear classifier in the high dimensional original space. In both case studies, the results showed that the DFA approach provides an intuitive interpretation of the differences between the groups, highlighting and reconstructing the most important statistical changes between the sample groups analyzed.

As future works, we can extend the DFA approach to several classes because both MLDA and SVM statistical learning methods used in this work can be generalized to multi-class problems. Additionally, we can perform similar experiments for a general separating hypersurface, that is, a non-linear discriminant features analysis. In this case, the normal direction changes when we travel along the separating boundary, which may bring new aspects for the reconstruction process.

Acknowledgments

The authors would like to thank the reviewers for their several comments that significantly improved this work. The authors also would like to thank Leo Leonel de Oliveira Junior for acquiring and normalizing the FEI face database under the grant FEI-PBIC 32-05. The SVM code used in this work is based on the quadratic programming solution created by Alex J. Smola. In addition, the authors would like to thank the support provided by PCI-LNCC, FAPESP (grant 2005/02899-4), and CNPq (grants 473219/04-2 and 472386/2007-7).

References

[1] R. Beale and T. Jackson. *Neural Computing*. MIT Press, 1994.

[2] C. J. C. Burges. A tutorial on support vector ma-

chines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[3] W. Chang. On using principal components before separating a mixture of two multivariate normal distributions. *Appl. Statist.*, 32(3):267–275, 1983.

[4] L. Chen, H. Liao, M. Ko, J. Lin, and G. Yu. A new lda-based face recognition system which can solve the small sample size problem. *Paterns Recognition*, 33:1713–1726, 2000.

[5] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *ECCV'98*, pages 484–498, 1998.

[6] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models- their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[7] T. F. Cootes, K.N. Walker, and C.J. Taylor. View-based active appearance models. In *4th International Conference on Automatic Face and Gesture Recognition*, pages 227–232, 2000.

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

[9] T. Fawcett. An introduction to roc analysis. *Pattern Recogn. Lett.*, 27(8):861–874, 2006.

[10] K. Fukunaga. Introduction to statistical pattern recognition. *Boston: Academic Press*, second edition, 1990.

[11] A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007.

[12] P. Golland, W. Grimson, M. Shenton, and R. Kikinis. Detection and analysis of statistical differences in anatomical shape. *Medical Image Analysis*, 9:69–86, 2005.

[13] P. Golland, W. Eric L. Grimson, Martha E. Shenton, and Ron Kikinis. Deformation analysis for shape based classification. *Lecture Notes in Computer Science*, 2082, 2001.

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

[15] C. J. Huberty. *Applied Discriminant Analysis*. John Wiley & Sons, INC., 1994.

[16] I. T. Jolliffe, B. J. T. Morgan, and P. J. Young. A simulation study of the use of principal component in linear discriminant analysis. *Journal of Statistical Computing*, 55:353–366, 1996.

[17] P. S. Rodrigues, G. A. Giraldi, Ruey-Feng Chang, and J. S. Suri. Non-extensive entropy for cad systems of breast cancer images. In *In Proc. of International Symposium on Computer Graphics, Image Processing and Vision - SIBGRAPI'06*, Manaus, Amazonas, Brazil, 2006.

[18] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC, 1986.

[19] J. S. Suri and R. M. Ragayyan. *Recent Advances in Breast Imaging, Mammography and Computer Aided Diagnosis of Breast Cancer*. SPIE Press, April 2006.

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

[21] C. E. Thomaz, N. A. O. Aguiar, S. H. A. Oliveira, F. L. S. Duran, G. F. Busatto, D. F. Gillies, and D. Rueckert. Extracting discriminative information from medical images: A multivariate linear approach. In *SIBGRAPI'06, IEEE CS Press*, pages 113–120, 2006.

[22] C. E. Thomaz, J. P. Boardman, S. Counsell, D.L.G. Hill, J. V. Hajnal, A. D. Edwards, M. A. Rutherford, D. F. Gillies, and D. Rueckert. A whole brain morphometric analysis of changes associated with preterm birth. In *SPIE International Symposium on Medical Imaging: Image Processing*, volume 6144, pages 1903–1910, 2006.

[23] C. E. Thomaz, J. P. Boardman, S. Counsell, D.L.G. Hill, J. V. Hajnal, A. D. Edwards, M. A. Rutherford, D. F. Gillies, and D. Rueckert. A multivariate statistical analysis of the developing human brain in preterm infants. *Image and Vision Computing*, 25(6):981–994, 2007.

[24] C. E. Thomaz, J. P. Boardman, D. L. G. Hill, J. V. Hajnal, D. D. Edwards, M. A. Rutherford, D. F. Gillies, and D. Rueckert. Using a maximum uncertainty lda-based approach to classify and analyse mr brain images. In *International Conference on Medical Image Computing and Computer Assisted Intervention MICCAI04*, pages 291–300, 2004.

[25] C. E. Thomaz and D. F. Gillies. A maximum uncertainty lda-based approach for limited sample size problems - with application to face recognition. In *SIBGRAPI'05, IEEE CS Press*, pages 89–96, 2005.

[26] C. E. Thomaz, D. F. Gillies, and R. Q. Feitosa. A new covariance estimate for bayesian classifiers in biometric recognition. *IEEE Transactions on*

Circuits and Systems for Video Technology, Special Issue on Image- and Video-Based Biometrics, 14(2):214–223, 2004.

[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 (JBCS)*, 12(2):7–18, 2006.

[28] C. E. Thomaz, P. S. Rodrigues, and G. A. Giraldi. Using face images to investigate the differences between lda and svm separating hyper-planes. In *II Workshop de Visao Computacional*, 2006.

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

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