# A Statistical Discriminant Model for Face Interpretation and Reconstruction

Edson C. Kitani[1], Carlos E. Thomaz[1], and Duncan F. Gillies[2]
[1]*Department of Electrical Engineering, Centro Universitário da FEI, São Paulo, Brazil*
[2]*Department of Computing, Imperial College, London, UK*
[1]{ekitani,cet}@fei.edu.br, [2]d.gillies@imperial.ac.uk

## Abstract

*Multivariate statistical approaches have played an important role of recognising face images and characterizing their differences. In this paper, we introduce the idea of using a two-stage separating hyper-plane, here called Statistical Discriminant Model (SDM), to interpret and reconstruct face images. Analogously to the well-known Active Appearance Model proposed by Cootes et. al, SDM requires a previous alignment of all the images to a common template to minimise variations that are not necessarily related to differences between the faces. However, instead of using landmarks or annotations on the images, SDM is based on the idea of using PCA to reduce the dimensionality of the original images and a maximum uncertainty linear classifier (MLDA) to characterise the most discriminant changes between the groups of images. The experimental results based on frontal face images indicate that the SDM approach provides an intuitive interpretation of the differences between groups, reconstructing characteristics that are very subjective in human beings, such as beauty and happiness.*

## 1. Introduction

The most successful statistical models for visual interpretation of face images have been based on Principal Component Analysis (PCA) [1, 3, 10]. These approaches have used as features either shapes [3] or textures [10] alone, or a combination of both [1]. Unfortunately, however, even in the PCA approach based on a combination of features, the sources of the shapes' and textures' variations have to be isolated in order to extract and interpret the most expressive differences in the training samples. For instance, in the well-known Active Appearance Model proposed by Cootes et. al. [1] the shape model is dissociate from the texture model and a manual annotation of landmarks is necessary to perform the statistical analysis.

In this paper, we introduce the idea of using a two-stage separating hyper-plane, here called Statistical Discriminant Model (SDM), to interpret and reconstruct face images. Analogously to the Cootes et al. approaches [1 – 4], SDM requires a previous alignment of all the images to a common template to minimise variations that are not necessarily related to differences between the faces. However, instead of using landmarks or annotations on the images, SDM is based on the idea of using PCA to reduce the dimensionality of the original images and a maximum uncertainty linear classifier (MLDA) [8] to characterise the most discriminant differences between the samples of images.

The remainder of this paper is divided as follows. In section 2, we briefly review PCA and highlight its importance on reducing the high dimensionality of face images. Section 3 describes the standard linear discriminant analysis (LDA) and states the reasons for using a maximum uncertainty version of this approach to perform the face experiments required. The estimation of the separating hyper-plane and the implementation of the Statistical Discriminant Model are described in Section 4. In section 5, we present experimental results of the PCA and SDM approaches on a face database maintained by the Department of Electrical Engineering at FEI. This section includes reconstruction experiments of face images using the SDM approach proposed. In the last section, section 6, the paper concludes with a short summary of the findings of this study and future directions.

## 2. Principal Component Analysis (PCA)

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. It is a well-known statistical technique that has been used in several image recognition problems, especially for dimensionality reduction. A comprehensive description of this multivariate statistical analysis method can be found in [6].

Let us consider the face recognition problem as an example to illustrate the main idea of the PCA. In any

image recognition, and particularly in face recognition, an input image with $n$ pixels can be treated as a point in an $n$-dimensional space called the image space. The coordinates of this point represent the values of each pixel of the image and form a vector $x^T = [x_1, x_2, \ldots, x_n]$ obtained by concatenating the rows (or columns) of the image matrix. 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, an input image with $n$ pixels can be projected onto a lower dimensional space without significant loss of information.

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

$$P^T S P = \Lambda . \quad \textbf{(1)}$$

It is a proven result that the set of $m$ ($m \leq n$) eigenvectors of $S$, which corresponds to the $m$ largest eigenvalues, minimises the mean square reconstruction error over all choices of $m$ orthonormal basis vectors [6]. Such a set of eigenvectors that defines a new uncorrelated coordinate system for the training set matrix $X$ is known as the principal components. In the context of face recognition, those $P_{pca}$ components are frequently called eigenfaces [10].

Therefore, although $n$ variables are required to reproduce the total variability (or information) of the sample $X$, much of this variability can be accounted for by a smaller number $m$ of principal components. That is, the $m$ principal components can then replace the initial $n$ variables and the original data set, consisting of $N$ measurements on $n$ variables, is reduced to a data set consisting of $N$ measurements on $m$ principal components.

## 3. Maximum Uncertainty LDA (MLDA)

The primary purpose of the Linear Discriminant Analysis, or simply LDA, is to separate samples of distinct groups by maximising their between-class separability while minimising their within-class variability. Although LDA does not assume that the populations of the distinct groups are normally distributed, it assumes implicitly that the true covariance matrices of each class are equal because the same within-class scatter matrix is used for all the classes considered.

Let the between-class scatter matrix $S_b$ be defined as

$$S_b = \sum_{i=1}^{g} N_i (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T \quad \textbf{(2)}$$

and the within-class scatter matrix $S_w$ be defined as

$$S_w = \sum_{i=1}^{g} (N_i - 1) S_i = \sum_{i=1}^{g} \sum_{j=1}^{N_i} (x_{i,j} - \bar{x}_i)(x_{i,j} - \bar{x}_i)^T \quad \textbf{(3)}$$

where $x_{i,j}$ is the $n$-dimensional pattern $j$ from class $\pi_i$, $N_i$ is the number of training patterns from class $\pi_i$, and $g$ is the total number of classes or groups. The vector $\bar{x}_i$ and matrix $S_i$ are respectively the unbiased sample mean and sample covariance matrix of class $\pi_i$ [6]. The grand mean vector $\bar{x}$ is given by

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{g} N_i \bar{x}_i = \frac{1}{N} \sum_{i=1}^{g} \sum_{j=1}^{N_i} x_{i,j} , \quad \textbf{(4)}$$

where $N$ is the total number of samples, that is, $N = N_1 + N_2 + \cdots + N_g$. It is important to note that the within-class scatter matrix $S_w$ defined in equation (3) is essentially the standard pooled covariance matrix $S_p$ multiplied by the scalar $(N - g)$, where $S_p$ can be written as

$$S_p = \frac{1}{N-g} \sum_{i=1}^{g} (N_i - 1) S_i = \frac{(N_1 - 1)S_1 + (N_2 - 1)S_2 + \cdots + (N_g - 1)S_g}{N - g} . \quad \textbf{(5)}$$

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

$$P_{lda} = \arg\max_{P} \frac{\left| P^T S_b P \right|}{\left| P^T S_w P \right|} . \quad \textbf{(6)}$$

The Fisher's criterion described in equation (6) is maximised when the projection matrix $P_{lda}$ is composed of the eigenvectors of $S_w^{-1} S_b$ with at most $(g - 1)$ nonzero corresponding eigenvalues [5]. This is the standard LDA procedure.

However, the performance of the standard LDA can

be seriously degraded if there is only a limited number of total training observations $N$ compared to the dimension of the feature space $n$. Since the within-class scatter matrix $S_w$ is a function of $(N-g)$ or less linearly independent vectors, its rank is $(N-g)$ or less. Therefore, $S_w$ is a singular matrix if $N$ is less than $(n+g)$, or, analogously, might be unstable if $N$ is not at least five to ten times $(n+g)$ [7].

To avoid the aforementioned critical issues of the standard LDA in limited sample and high dimensional problems, we have calculated $P_{lda}$ by using a maximum uncertainty LDA-based approach (MLDA) that considers the issue of stabilising the $S_w$ estimate with a multiple of the identity matrix [8, 9]. In a previous study [8] with application to the face recognition problem, Thomaz and Gillies showed that the MLDA approach improved the LDA classification performance with or without a PCA intermediate step and using less linear discriminant features [8].

The MLDA algorithm can be described as follows:

i. Find the $\Phi$ eigenvectors and $\Lambda$ eigenvalues of $S_p$, where $S_p = S_w / [N-g]$;

ii. Calculate the $S_p$ average eigenvalue $\overline{\lambda}$, that is,

$$\overline{\lambda} = \frac{1}{n}\sum_{j=1}^{n}\lambda_j = \frac{trace(S_p)}{n}; \qquad \textbf{(7a)}$$

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

$$\Lambda^* = diag[\max(\lambda_1, \overline{\lambda}), ..., \max(\lambda_n, \overline{\lambda})]; \qquad \textbf{(7b)}$$

iv. Form the modified within-class scatter matrix

$$S_w^* = S_p^*(N-g) = (\Phi \Lambda^* \Phi^T)(N-g). \qquad \textbf{(7c)}$$

The maximum uncertainty LDA (MLDA) is constructed by replacing $S_w$ with $S_w^*$ in the Fisher's criterion formula described in equation (6). It is based on the idea [8] 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 minimising a more difficult but appropriate "inflated" within-class scatter matrix would also minimise a less reliable "shrivelled" within-class estimate.

## 4. Statistical Discriminant Model (SDM)

The Statistical Discriminant Model proposed in this work is essentially a two-stage PCA+MLDA linear classifier that reduces the dimensionality of the original images and extracts discriminant information from images.

In order to estimate the SDM separating hyperplane, we use training examples and their corresponding labels to construct the classifier. First a training set is selected and the average image vector of all the training images is calculated and subtracted from each $n$-dimensional vector. Then the training matrix composed of zero mean image vectors is used as input to compute the PCA transformation matrix. The columns of this $n$ x $m$ transformation matrix are eigenvectors, not necessarily in eigenvalues descending order. We have retained all the PCA eigenvectors with non-zero eigenvalues, that is, $m = N-1$, to reproduce the total variability of the samples with no loss of information. The zero mean image vectors are projected on the principal components and reduced to $m$-dimensional vectors representing the most expressive features of each one of the $n$-dimensional image vector. Afterwards, this $N$ x $m$ data matrix is used as input to calculate the MLDA discriminant eigenvector, as described in the previous section. Since in this work we have limited ourselves to two-group classification problems, there is only one MLDA discriminant eigenvector. The most discriminant feature of each one of the $m$-dimensional vectors is obtained by multiplying the $N$ x $m$ most expressive features matrix by the $m$ x 1 MLDA linear discriminant eigenvector. Thus, the initial training set of face images consisting of $N$ measurements on $n$ variables, is reduced to a data set consisting of $N$ measurements on only 1 most discriminant feature.

Once the two-stage SDM classifier has been constructed, we can move along its corresponding projection vector and extract the discriminant differences captured by the classifier. Any point on the discriminant feature space can be converted to its corresponding $n$-dimensional image vector by simply: (1) multiplying that particular point by the transpose of the corresponding linear discriminant vector previously computed; (2) multiplying its $m$ most expressive features by the transpose of the principal components matrix; and (3) adding the average image calculated in the training stage to the $n$-dimensional image vector. Therefore, assuming that the spreads of the classes follow a Gaussian distribution and applying limits to the variance of each group, such as $\pm 2$ sd, where sd is the standard deviation of each group, we can move along the SDM most discriminant features and map the results back into the image domain.

(a)

(b)

Figure 1. Samples of the female versus male (a) and non-smiling versus smiling training sets (b).

## 5. Experimental Results

We have used frontal images of a face database maintained by the Department of Electrical Engineering of FEI to carry out the experiments. This database contains a set of face images taken between June 2005 and March 2006 at the Artificial Intelligence Laboratory in São Bernardo do Campo, with 14 images for each of 118 individuals – a total of 1652 images*. All images are colourful 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.

To minimise image variations that are not necessarily related to differences between the faces, we aligned first 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 64x64 pixels and converted to 8-bit grey scale.

We have carried the following two-group statistical analyses: female versus male experiments, and non-smiling versus smiling experiments. The idea of the first discriminant experiment is to evaluate the statistical approaches on a discriminant task where the differences between the groups are evident. In contrast, the second experiment, i.e. non-smiling versus smiling samples, poses an alternative analysis where there are subtle differences between the groups. Since the number of female images is limited and equal to 49, we have composed the female/male training set of 49 frontal female images and 49 frontal male images. For the smiling/non-smiling experiments, we have used the 49 frontal male images previously selected and their corresponding frontal smiling images. All faces are mainly represented by subjects between 19 and 30 years old with distinct appearance, hairstyle, and adorns. Figure 1 shows some examples of these two training sets selected.

### 5.1. PCA Results

In this section, we describe the most expressive features captured by PCA. As the average face image is an $n$-dimensional point ($n = 4096$) 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. [1 – 4], we have reconstructed the new average face images by changing each principal component separately using the limits of $\pm\sqrt{\lambda_i}$, where $\lambda_i$ are the corresponding largest eigenvalues.

Figure 2 illustrates these transformations on the first three most expressive principal components using the female/male training set. As 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 principal component (middle), in turn, models variations related to the grey-level of the faces and hair, but it is not clear which specific variation this component is actually capturing. The last principal component considered, the third component (bottom), models mainly the size of the head of the training samples. It is important to note that as the female/male training set has a very clear separation be-

* All these images are available upon request (cet@fei.edu.br).

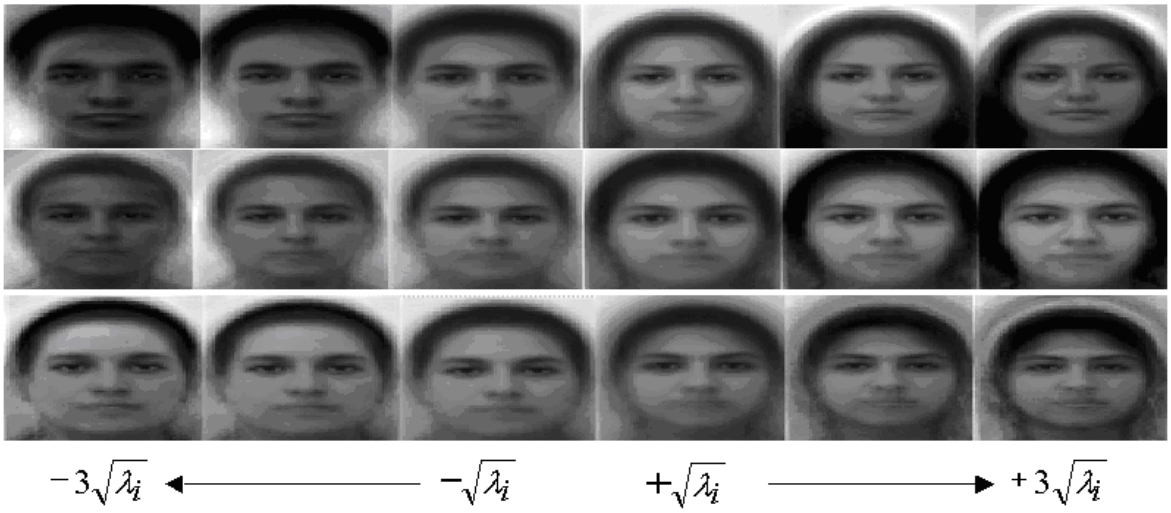


Figure 2. PCA results using the female/male training set.

tween the groups, the principal components have kept this separation and when we move along each principal component axis we can see this major difference between the samples, even though subtly, such as in the third principal component illustrated.

Figure 3 presents the three most expressive variations captured by PCA using the non-smiling/smiling training set, which is composed of male images only. Analogously to the female/male experiments, the first principal component (on the top) captures essentially the changes in illumination, the second principal component (middle) models variations particularly in the head shape, and the third component (bottom) captures variations in the facial expression among others.

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 can be seen in Figure 3, although the third principal component (bottom) models some facial expression variation, this specific variation has been captured by other principal components as well including other image artefacts. Likewise, as Figure 2 illustrates, although the first principal component (top) models gender variation, other changes have been modelled concurrently,

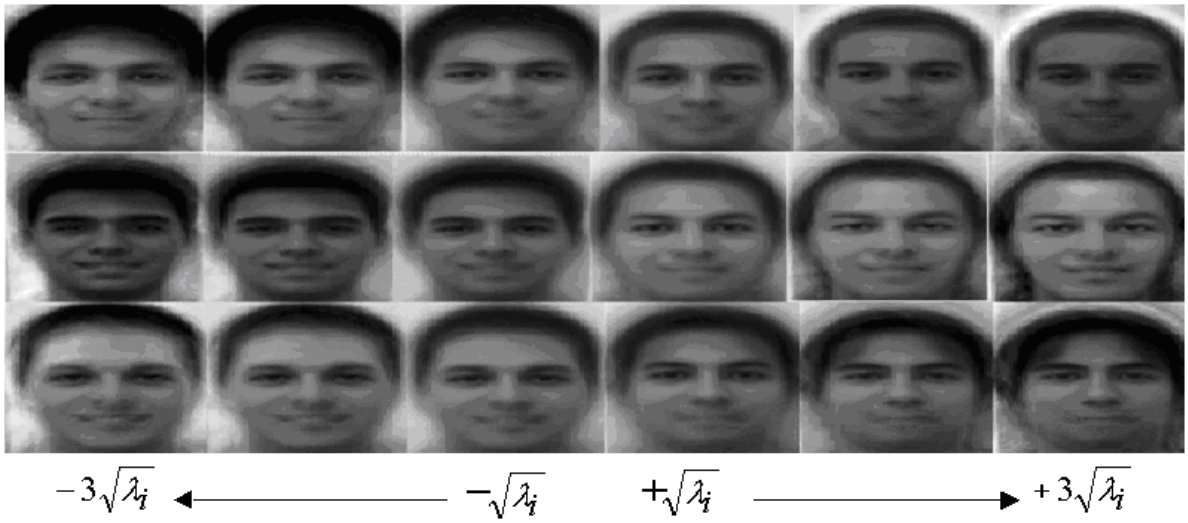


Figure 3. PCA results using the non-smiling/smiling training set (male images only).

-2sd -1sd Mean +1sd +2sd

Figure 4. SDM results using the female/male training set.

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.

## 5.2. SDM Results

As described earlier, in order to estimate the SDM separating hyperplane, we have used the female/male and non-smiling/smiling training sets previously selected and their corresponding labels to construct the classifier. Since in these experiments we have limited ourselves to two-group classification problems, there is only one SDM discriminant eigenvector. Therefore, assuming that the spreads of the classes follow a Gaussian distribution and applying limits to the variance of each group, such as $\pm 2$ sd, where 'sd' is the standard deviation of each group, we can move along the SDM most discriminant features and map the results back into the image domain for visual analysis.

Figure 6 presents the SDM most discriminant features for the gender experiments. It displays the image regions captured by the SDM approach that change when we move from one side (left, male) of the dividing hyper-plane to the other (right, female), following limits to the standard deviation ($\pm 2$ sd) of each sample group. As can be seen, the SDM hyper-plane effectively extracts the 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 artefacts.

Figure 5 shows the SDM most discriminant features for the facial expression experiments. Analogously to the gender experiments, Figure 5 displays the image regions captured by the SDM classifier that change when we move from one side (left, smiling) of the dividing hyper-plane to the other (right, non-smiling), following limits to the standard deviation ($\pm 2$ sd) of each sample group. As can be seen, the SDM hyper-plane effectively extracts the group differences, showing exactly what we should expect intuitively from a face image when someone changes their expression from smiling to non-smiling. In fact, it is possible to note that the SDM most discriminant direction has predicted a facial expression not necessarily present in our corresponding smiling/non-smiling training set, that is, the "definitely non-smiling" or may be "anger" status represented by the image +2sd in Figure 5.

Analogously to the PCA experiments, all SDM reconstructions have been made using the average face image of the corresponding training sets. However, it is possible to project any face image on the SDM feature space, move along its corresponding most discriminant features, and map the changes back to the original image space. Figure 6 shows these experimental results when we move an example image along the male/female (Figure 6a) and smiling/non-smiling (Figure 6b) hyper-planes previously calculated. As can be seen in Figure 6a, the most discriminant features be-

-2sd -1sd Mean +1sd +2sd

Figure 5. SDM results using the smiling/non-smiling training set.

(a)

(b)

Figure 6. SDM results when we move an example image along the male/female (a) and smiling/non-smiling (b) hyper-planes.

tween a male and female face images have been incorporated on the example image when we move it to the male side of the dividing hyper-plane, such as the thickening of the lips, nose, and eyebrows. In contrast, since the example chosen is from a woman, almost no facial changes occurs when we move the same example to the other side of the hyper-plane, that is, to the female side. Also, according to Figure 6b, it is possible to see that the SDM linear classifier has incorporated all the most discriminant facial changes that we intuitively expect when we change our facial expression from smiling to non-smiling status. It is important to note in this case where most of the facial changes are localised around the mouth that only the differences related to the facial expression differences have changed on the image with no impact on other face features, such as hair-style, forehead, eyebrows, and chin.

## 6. Conclusion

In this work, we introduced the idea of using the PCA+MLDA two-stage linear classifier to interpret and reconstruct frontal face images rather than recognising subjects. Differently from other statistical approaches, our method is based on a supervised separation between the whole images and not on the use of landmarks and isolated models for the shapes' and textures' variations. The experiments carried out in this work showed that subjective information such as beauty and happiness can be efficiently captured by a linear classifier when we pre-process the face images using a simple affine transformation. The results presented in this paper suggested that the statistical discriminant model proposed could be useful to reconstruct not only frontal face images but also face images with different profiles. Further work is being undertaken to investigate this possibility.

## Acknowledgments

The authors would like to thank Leo Leonel de Oliveira Junior for acquiring and normalizing the FEI database under the grant FEI-PBIC 32-05.

## References

[1] T. F. Cootes, G. J. Edwards, C. J. Taylor, "Active Appearance Models", H.Burkhardt and B. Newmann editors, in Proceedings of ECCV'98, vol. 2, pp. 484-498, 1998.

[2] T.F Cootes, A. Lanitis, "Statistical Models of Appearance for Computer Vision", Technical report, University of Manchester, 125 pages, 2004.

[3] T. F. Cootes, C. J. Taylor, D. H. Cooper, J. Graham, "Active Shape Models- Their Training and Application", *Computer Vision and Image Understanding*, vol. 61, no.1, pp. 38-59, 1995.

[4] T.F. Cootes, K.N. Walker, C.J. Taylor, "View-Based Active Appearance Models", In 4th International Conference on Automatic Face and Gesture Recognition, Grenoble, France, pp. 227-232, 2000.

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

[6] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, second edition. Boston: Academic Press, 1990.

[7] A. K. Jain and B. Chandrasekaran, "Dimensionality and Sample Size Considerations in Pattern Recognition Practice", *Handbook of Statistics*, 2, pp. 835-855, 1982.

[8] C. E. Thomaz and D. F. Gillies, "A Maximum Uncertainty LDA-based approach for Limited Sample Size problems - with application to Face Recognition", in Proceedings of SIBGRAPI'05, IEEE CS Press, pp. 89-96, 2005.

[9] 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*, vol. 14, no. 2, pp. 214-223, 2004.

[10] M. Turk, A. Pentland, Eigenfaces for Recognition, *Journal of Cognitive Neuroscience*, MIT, vol. 73, pp. 71-86, 1991.