Feature Extraction for Nearest Neighbor Classification: Application to Gender Recognition

David Masip,* Jordi Vitrià†
Centre de Visió per Computador (CVC), Departemente Informàtica,
Universitat Autònoma de Barcelona, Bellaterra 08193, Spain

In this article, we perform an extended analysis of different face-processing techniques for gender recognition problems. Prior research works show that support vector machines (SVM) achieve the best classification results. We will show that a nearest neighbor classification approach can reach a similar performance or improve the SVM results, given an adequate selection of features of the input data. This selection is performed using a dimensionality reduction technique based on a modification of nonparametric discriminant analysis, designed to improve the nearest neighbor classification. The choice of nearest neighbor is especially justified by the use of a large database. We also analyze a nonlinear algorithm, locally linear embedding, and its supervised version. Given that this technique is focused on preserving the local configuration of the neighborhood of each point, it should be a priori a good dimensionality reduction technique for extracting good features for nearest neighbor classification. A complete comparative study with the most classical face-processing techniques is also performed. © 2005 Wiley Periodicals, Inc.

1. INTRODUCTION

In the last few years, computational resources have become cheaper, smaller, and more powerful. This evolution will allow the progressive introduction of technology in our everyday lives and new applications dealing with cameras will emerge. Some of the most important are related to face-classification techniques. Typical examples are face recognition applied to security systems, face verification in authentication schemes, face and gesture analysis in user friendly interfaces, and gender and ethnicity recognition for applications of reactive marketing. In this article, we will deal with a gender recognition problem. We will show different schemes to solve it, and the results can be taken into account as a benchmark of techniques when we need to solve more general face classification tasks.

*Author to whom all correspondence should be addressed: e-mail: davidm@cvc.uab.es.
†e-mail: jordi@cvc.uab.es.

INTERNATIONAL JOURNAL OF INTELLIGENT SYSTEMS, VOL. 20, 561–576 (2005)
© 2005 Wiley Periodicals, Inc. Published online in Wiley InterScience
(www.interscience.wiley.com). • DOI 10.1002/int.20081

We propose a nearest neighbor (NN) approach for gender recognition. This election can be justified by the use of large face databases and by taking advantage of proper feature extraction techniques. We will evaluate different feature extraction techniques applied to NN classification of gender in human faces, showing that the nearest neighbor approach can compete with the best techniques for this task, such as Principal Component Analysis (PCA)[1] and Support Vector Machines.[2] To extract the optimal features for nearest neighbor classification we will evaluate linear and nonlinear techniques, and we will also show that the adequate modification of a discriminant analysis algorithm can provide the best features for nearest neighbor classification. The main goal of this modification is to minimize the intraclass variations while the extraclass variance is maximized.

Data dimensionality reduction techniques have often been used in pattern recognition, and have been applied to different objectives. One possible application is to use these techniques for data compression, reducing the amount of data of each sample by projecting the sample in a reduced space. Another application is the improvement of the classification algorithms by selecting the features that best separate the different classes or by simply reducing the noise present in natural images. Sometimes the reduction of dimensionality helps the training of the classifier too, by reducing the number of parameters to estimate.

Perhaps the most popular technique applied to face classification is Principal Component Analysis and eigenfaces.[3,4] The goal in PCA is to find a linear projection to a low dimensional subspace, trying to preserve the maximum amount of variance of the input data. Other criteria can be applied to find the optimal projection, such as statistical independence (Independent Component Analysis[5]) and nonnegativity (Nonnegative Matrix Factorization[6]). In the next section we will give an overview of the state of the art of gender recognition.

If we take into account the labels of the training samples in the dimensionality reduction process, we can find the most discriminative features in a reduced space, where the distance between different class samples is maximized. The classic discriminative technique is Fisher Linear Discriminant (FLD).[7] But FLD has two important drawbacks: The resulting dimensionality is upper bounded by the number of classes, and this complicates its application to the gender recognition problem (where there are only two classes). FLD also assumes Gaussian densities distributions in sample data, which degrades the performance in the case of more general distributions. In Section 3 we will introduce the Nonparametric Discriminant Analysis[8] and our modification of the original algorithm to overcome this drawbacks.

Another approach to dimensionality reduction is the use of nonlinear techniques. One of them is Locally Linear Embedding (LLE),[9] which will be analyzed in Section 4. The goal of LLE is to find the low dimensional space that best preserves the local configuration of each point with respect to its nearest neighbors, so it could be expected that the use of this technique would be useful for a nearest neighbor classification. We also evaluate a simple supervised version of the algorithm,[10] which improves the results considerably.

In Section 4 we will show some preprocessing and postprocessing steps to improve the NN classification, reducing the noise present in the data samples and using a combination of classifiers. In Section 5 we will show the face database and the gender recognition experiments performed. The final conclusions are shown in the last section.

2. PREVIOUS WORKS IN GENDER RECOGNITION

Humans are able to distinguish the gender from faces' images with high accuracy. In fact, some psychological studies[11] have shown that we are able to achieve accuracies close to 96%, using images where the additional information of the hair has been eliminated. It would be interesting to find computational gender classifiers that could reach similar ratios. Computational face recognition techniques have been the subject of intensive research during the last few years; nevertheless few algorithms have been proposed in the field of gender recognition. Here we will highlight some of the most important ones.

Different approaches have been used in the literature, which can be divided into two groups: geometry-based and appearance-based classification techniques. In the first case, a set of features extracted from each image is used to train a classifier, for example, some kind of distance (between eyes, eyebrows, etc.), or size (face, mouthm and nose size, etc.). Brunelli and Poggio[12] used a set of 16 geometric features per image to train two hyper basis function networks, and achieved accuracies of 79% in a database composed of 168 training images. In another work by Burton et al.,[11] discriminant analysis was used over a set of 73 features (such as distances between key points, ratios and angles formed by the key points, etc.), achieving an 85% accuracy.

On the other hand, in the appearance-based models the classifier is trained using the whole image instead of using some geometric extracted features. In an experiment performed in Burton et al.,[13] human subjects were asked to identify the gender of a set of pictures of faces and a set of three-dimensional (3-D) laser-scanned representations of the same faces. The results showed that it was more difficult to discriminate between classes in the 3-D images, which suggests that features like global skin texture are very important in the gender recognition process. In a similar way another experiment was performed using face pictures and inverted pictures.[14] The accuracy in the inverted pictures decreased significantly. Perhaps the most representative appearance-based method is the eigenface approach. Abdi et al.[15] trained a perceptron classifier using PCA-based features of the input images, achieving a performance of the 91.8%.

Cottrell[16] used a two-layer neural network approach, where each face image was compressed in the first layer of the network and classified in the second layer. They obtained an accuracy of 63% using only 64 training images. In a similar work by Golomb et al.[17] a system named SEXNET was used with a 91.9% accuracy. They used a neural network with 40 units to encode (compress) the 900-dimensional face image, and then they used two layers of 40 hidden units to classify the encodings. Tamura et al.[18] also used a neural network to identify sex, achieving accuracies close to 90% even using reduced 8×8 central face images.

Gutta et al.[19] proposed a hybrid approach using radial basis function (RBF) networks and inductive decision trees achieving an accuracy of 96%. Moghaddam and Yang[2] obtained the best performance, achieving 96.6% using a large face database (1755 faces), using SVM with RBF kernels.

3. DISCRIMINANT ANALYSIS

The most simple classification rule is the one implemented by the NN classifier. As we use the NN classification rule, a proper feature extraction algorithm can simplify the classification step, and Discriminant Analysis can be very useful for this task. In this section we will introduce classic Fisher Discriminant Analysis, to show later how its nonparametric extension can solve the main drawbacks of FLD: Gaussian distribution assumption and reduced dimensionality of the generated subspaces. We will also show our modification of classic nonparametric discriminant analysis (NDA),[8] which is expected to improve the NN classification.

3.1. Fisher Discriminant Analysis

The goal of discriminant analysis is to find the features that best separate the different classes. One of the most used criterions $\mathcal{J}$ to reach it is to maximize

$$\mathcal{J} = \operatorname{tr}(\boldsymbol{S}^E \boldsymbol{S}^I) \tag{1}$$

where the matrices $\boldsymbol{S}^E$ and $\boldsymbol{S}^I$ generally represent the scatter of sample vectors between different classes and within a class, respectively. It has been shown (see Refs. 20 and 21) that the $M \times D$ linear transform that satisfies

$$\hat{\boldsymbol{W}} = \arg \max_{\boldsymbol{W}^T \boldsymbol{S}^I \boldsymbol{W} = \boldsymbol{I}} \operatorname{tr}(\boldsymbol{W}^T \boldsymbol{S}^E \boldsymbol{W}) \tag{2}$$

optimizes the separability measure $\mathcal{J}$. This problem has an analytical solution based on the eigenvectors of the scatter matrices. The algorithm presented in Table I

Table I. General algorithm for solving the discriminability optimization problem stated in Equation (2).

1. Given $\boldsymbol{X}$ the matrix containing data samples placed as N D-dimensional columns, $\boldsymbol{S}^I$ the within class scatter matrix, and M maximum dimension of discriminant space.
2. Compute eigenvectors and eigenvalues for $\boldsymbol{S}^I$. Make $\boldsymbol{\Phi}$ the matrix with the eigenvectors placed as columns and $\boldsymbol{\Lambda}$ the diagonal matrix with only the nonzero eigenvalues in the diagonal. M^I is the number of nonzero eigenvalues.
3. Whiten the data with respect to $\boldsymbol{S}^I$, to obtain M^I-dimensional whitened data,

$$\boldsymbol{Z} = \boldsymbol{\Lambda}^{-1/2} \boldsymbol{\Phi}^T \boldsymbol{X}$$

4. Compute $\boldsymbol{S}^{\mathrm{E}}$ on the whitened data.
5. Compute eigenvectors and eigenvalues for $\boldsymbol{S}^E$ and make $\boldsymbol{\Psi}$ the matrix with the eigenvectors placed as columns and sorted by decreasing eigenvalue value.
6. Preserve only the first $M^E = \min\{M^I, M, \operatorname{rank}(\boldsymbol{S}^E)\}$ columns, $\boldsymbol{\Psi}_M = \{\psi_1, \ldots, \psi_{M^E}\}$ (those corresponding to the M^E largest eigenvalues).
7. The resulting optimal transformation is $\hat{\boldsymbol{W}} = \boldsymbol{\Psi}_M^T \boldsymbol{\Lambda}^{-1/2} \boldsymbol{\Phi}^T$ and the projected data, $\boldsymbol{Y} = \hat{\boldsymbol{W}} \boldsymbol{X} = \boldsymbol{\Psi}_M^T \boldsymbol{Z}$.

obtains this solution.[21] The most widely spread approach for defining the within- and between-class scatter matrices is the one that makes use of only up to second order statistics of the data. This was done in a classic paper by Fisher[7] and the technique is referred to as Fisher Discriminant Analysis (FLD). In FLD the within-class scatter matrix is usually computed as a weighted sum of the class-conditional sample covariance matrices. If equiprobable priors are assumed for classes C_k, $k = 1, \dots, K$ then

$$\boldsymbol{S}^I = \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{\Sigma}_k \tag{3}$$

where $\boldsymbol{\Sigma}_k$ is the class-conditional covariance matrix, estimated from the sample set. The between-class scatter matrix is defined as

$$\boldsymbol{S}^E = \frac{1}{K} \sum_{k=1}^{K} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_k - \boldsymbol{\mu}_0)^T \tag{4}$$

where $\boldsymbol{\mu}_k$ is the class-conditional sample mean and $\boldsymbol{\mu}_0$ is the unconditional (global) sample mean.

Notice the rank of $\boldsymbol{S}^E$ is $K-1$, so the number of extracted features is, at most, one less than the number of classes. Also notice the parametric nature of the scatter matrix. The solution provided by FLD is blind beyond second-order statistics. So we cannot expect our method to accurately indicate which features should be extracted to preserve any complex classification structure.

3.2. Nonparametric Discriminant Analysis

In Ref. 8, Fukunaga and Mantock present a nonparametric method for discriminant analysis in an attempt to overcome the limitations present in FLD. In NDA the between-class scatter $\boldsymbol{S}^E$ is of a nonparametric nature. This scatter matrix is generally full rank, thus loosening the bound on extracted feature dimensionality. Also, the nonparametric structure of this matrix inherently leads to extracted features that preserve relevant structures for classification. We briefly expose this technique, extensively detailed in Ref. 21.

In NDA, the between-class scatter matrix is obtained from vectors locally pointing to another class. This is done as follows. The extraclass nearest neighbor for a sample $\boldsymbol{x} \in C_k$ is defined as $\boldsymbol{x}^E = \{\boldsymbol{x}' \in \bar{C}_k / \|\boldsymbol{x}' - \boldsymbol{x}\| \leq \|z - \boldsymbol{x}\|, \forall z \in \bar{C}_k\}$. In the same fashion we can define the set of intraclass nearest neighbors as $\boldsymbol{x}^I = \{\boldsymbol{x}' \in L_c / \|\boldsymbol{x}' - \boldsymbol{x}\| \leq \|z - \boldsymbol{x}\|, \forall z \in C_k\}$.

From these neighbors, the extraclass differences are defined as $\boldsymbol{\Delta}^E = \boldsymbol{x} - \boldsymbol{x}^E$ and the intraclass differences as $\boldsymbol{\Delta}^I = \boldsymbol{x} - \boldsymbol{x}^I$. Notice that $\boldsymbol{\Delta}^E$ points locally to the nearest class (or classes) that does not contain the sample. The nonparametric between-class scatter matrix is defined as (assuming uniform priors)

$$\boldsymbol{S}^E = \frac{1}{N} \sum_{n=1}^{N} (\boldsymbol{\Delta}_n^E)(\boldsymbol{\Delta}_n^E)^T \tag{5}$$

where $\boldsymbol{\Delta}_n^E$ is the extraclass difference for sample $\boldsymbol{x}_n$.

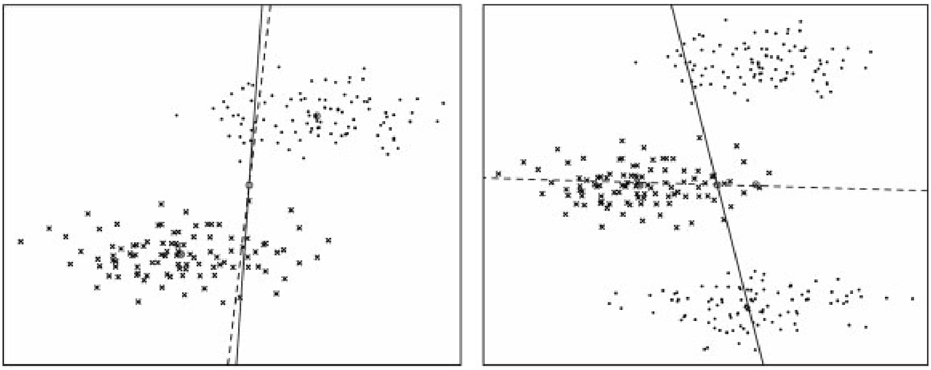


Figure 1. First directions of NDA (solid line) and FLD (dashed line) projections, for two artificial data sets. Observe the results in the right-hand figure, where the FLD assumptions are not met.

A parametric form is chosen for the within-class scatter matrix $\boldsymbol{S}^I$, defined as in Equation 3. Figure 1 illustrates the differences between NDA and FLD in two artificial data sets, one with Gaussian classes where results are similar, and one where FLD assumptions are not met. For the second case, the bimodality of one of the classes displaces the class mean, introducing errors in the estimate of the parametric version of $\boldsymbol{S}^E$. The nonparametric version is not affected by this situation.

We now make use of the introduced notation to examine the relationship between NN and NDA. This results in a modification of the within-class covariance matrix, which we also introduce.

Given a training sample $\boldsymbol{x}$, the accuracy of the 1-NN rule can be directly computed by examining the ratio $\|\boldsymbol{\Delta}^E\|/\|\boldsymbol{\Delta}^I\|$. If this ratio is more than one, $\boldsymbol{x}$ will be correctly classified. Given the $M \times D$ linear transform $\boldsymbol{W}$, the projected distances are defined as $\boldsymbol{\Delta}_W^{E,I} = \boldsymbol{W}\boldsymbol{\Delta}^{E,I}$. Notice that this definition does not exactly agree with the extra- and intraclass distances in projection space because, except for the orthonormal transformation case, we have no warranty on distance preservation. Equivalence of both definitions is asymptotically true. By the above remarks it is expected that optimization of the following objective function should improve or at least not downgrade NN performance:

$$\hat{W} = \arg \max_{E\{\|\boldsymbol{\Delta}_W^I\|^2\}=1} E\{\|\boldsymbol{\Delta}_W^E\|^2\} \tag{6}$$

This optimization problem can be interpreted as: Find the linear transform that maximizes the distance between classes, preserving the expected distance among the members of a single class. Considering that,

$$E\{\|\boldsymbol{\Delta}_W\|^2\} = E\{(\boldsymbol{W}\boldsymbol{\Delta})^T(\boldsymbol{W}\boldsymbol{\Delta})\} = \operatorname{tr}(\boldsymbol{W}^T\boldsymbol{\Delta}\boldsymbol{\Delta}^T\boldsymbol{W}) \tag{7}$$

where $\boldsymbol{\Delta}$ can be $\boldsymbol{\Delta}^I$ or $\boldsymbol{\Delta}^E$. Replacing Equation 7 in Equation 6 we have that this last equation is a particular case of Equation 2. Additionally, the formulas for the within- and between-class scatter matrices are directly extracted from this equation. In

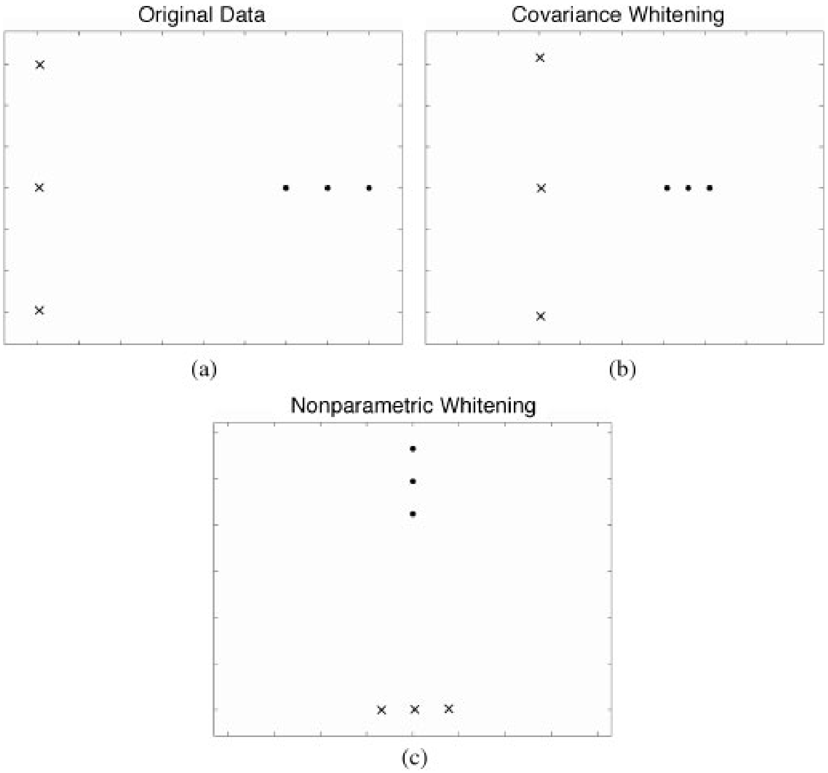


Figure 2. (a) Original data from a toy data set. (b) The same data whitened using the covariance matrix of classic FLD. (c) Whitened data using the nonparametric modification. The normalization in the last case is performed taking into account the distribution of the intraclass nearest neighbor distances.

this case, the between-class scatter matrix agrees with Equation 5, but the within-class scatter matrix is now defined in a nonparametric fashion:

$$S_w = \frac{1}{N} \sum_{n=1}^{N} \mathbf{\Delta}_n^I \mathbf{\Delta}_n^{I^T} \tag{8}$$

Given that we have an optimization problem of the form given in Equation 2, the algorithm presented in Table I can also be applied to the optimization of our proposed objective function 6. In Figure 2 a graphical example of the intra-class normalization is shown. Points of the same class are normalized according to the distances between each point and its nearest neighbor.

4. LOCALLY LINEAR EMBEDDING

As has been shown, NDA can be seen as a dimensionality reduction technique that is optimal for the nearest neighbor classification rule. Recently, a new

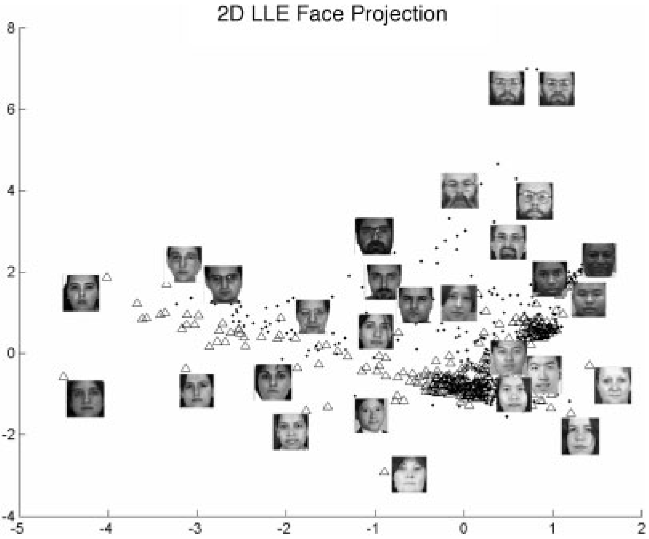


Figure 3. Two-dimensional reduction of faces using LLE. Original faces are plotted near each reduced point. Triangles stand for female subjects and dots for male subjects. As can be observed some characteristics such as global illumination, beard presence (on the top right corner), or ethnicity are captured by LLE embedding.

nonsupervised technique has been proposed that presents, in principle, the same property: Local Linear Embedding (LLE).[9] The goal of LLE is to find a mapping from a high dimensional space to a low dimensional one too, but performed in a nonlinear way. Sometimes the high dimensional data lie in a nonlinear manifold that can be represented using less dimensions than the dimensionality of the original points. To reach this objective, LLE takes into account the restriction that neighborhood points in the high dimensional space must remain in the same neighborhood in the low dimensional space, and placed in a similar relative spatial situation (it does not change the local structure of the nearest neighbors of each point).

Given N D-dimensional training images as input vectors $\boldsymbol{x}_n$ to the LLE technique, a three-step algorithm detailed in Table II is performed to find the low dimensional space. This low dimensional representation of the data preserves local neighborhoods, ensuring that the nearest neighbor classification rule will not degrade after this transformation. In Figure 3, we can see an example of a two-dimensional embedding of face image.

4.1. Supervised LLE

As we are dealing with classification algorithms, an interesting approach is to consider the class membership of the train vectors to achieve class separation.[10] The main difference between LLE and SLLE is the first step of the algorithm, the search of the nearest neighbors. Whereas LLE looks for the K nearest neighbors of

Table II. LLE algorithm.

1. First we compute the K nearest neighbors of each point.
2. Capture the local geometry of the input data, using a set of $\boldsymbol{W}$ coefficients per each point, corresponding to the weights W_{nk} that best reconstruct the vector $\boldsymbol{x}_n$ from its K nearest neighbors $\boldsymbol{x}_{n_k}$, minimizing the error reconstruction equation:

$$\varepsilon(\boldsymbol{W}) = \sum_{n=1}^{N} \left| \boldsymbol{x}_n - \sum_{k=1}^{K} W_{nk} \boldsymbol{x}_{n_k} \right|^2 \tag{9}$$

   To find the vectors that minimize this equation, a least-squares problem must be solved. For more details see Ref. 22.
3. In the last step the coordinates of each point in the low dimensional space $d' \ll D$ are computed as the vectors $\boldsymbol{y}_n$ that best minimize the equation

$$\theta(\boldsymbol{y}) = \sum \left| \boldsymbol{y}_i - \sum W_{nk} \boldsymbol{y}_{n_k} \right|^2 \tag{10}$$

   The weights found during the previous stage are constant, and we want to find the low dimensional outputs $\boldsymbol{y}_n$ that best reconstruct each vector using the information of these weights, which capture the local geometric properties of each point in the original space. A new sparse matrix $\boldsymbol{M}$ is created and defined as

$$M_{ij} = \delta_{ij} - W_{ij} - W_{ji} + \sum_{k=1}^{K} W_{ki} W_{kj} \tag{11}$$

   It can be proved that the output vectors $\boldsymbol{y}_n$ are the $d' + 1$ eigenvectors of the matrix $\boldsymbol{M}$ associated with the lowest eigenvalues (see Refs. 9 and 22 for more details).

each point from the whole data set, SLLE only searches the K nearest neighbors from the set of points belonging to the same class of each point. So the weights computed in the second step encode the best way of reconstructing each point from its nearest neighbors of the same class of the point. The rest of the algorithm is identical to the exposed LLE.

4.2. Projections of the Test Vectors

As has been shown, the LLE algorithm is a globally nonlinear technique. This property has some advantages when finding the underlying manifolds, but there is an important drawback when a new point $\boldsymbol{u}$ is entered as a new input to the system. An approximation of the mapping is necessary to avoid rerunning the algorithm each time (solving the expensive eigenvector problem). Parametric (probabilistic) and nonparametric models have been used to solve this problem (see Refs. 9 and 22). In Ref. 23 the use of a multilayer perceptron (MLP) neural network is proposed to learn the projection, and then it is only necessary to run a forward step in the MLP to find the coordinates of each new unseen input vector.

Another approach[24] is to the find the k nearest neighbors of the new point $\boldsymbol{u}$ using the points of the training set $\boldsymbol{x_n}$. Then compute the reconstructing weights $\boldsymbol{W_{nk}}$ using this neighbors. And finally compute the coordinates of the point $\boldsymbol{u}$ in the reduced space as

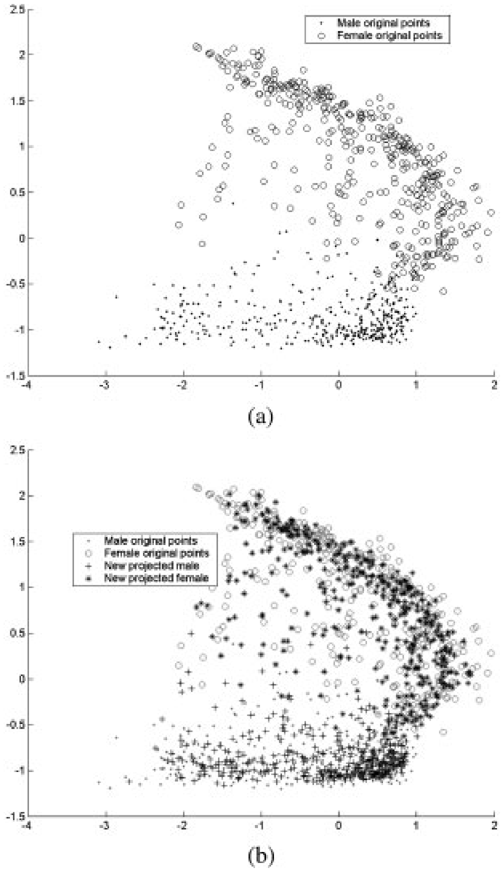


Figure 4. (a) Example of an embedding of two-class 500 data samples using LLE (to a two-dimensional space). (b) The projection of 500 new unseen data samples.

$$\boldsymbol{y} = \sum_{j=1}^{k} W_j \cdot \boldsymbol{y}_{\boldsymbol{N}(\boldsymbol{j})} \tag{12}$$

where $\boldsymbol{y}_{\boldsymbol{N}(\boldsymbol{j})}$ are the points in the low dimensional space corresponding to the used nearest neighbors of the point $\boldsymbol{u}$ in the high dimensional space. As can be seen in Figures 4 and 5, the projections obtained still preserve some class separability, especially when the supervised LLE algorithm is used. The confusion between classes is directly correlated to the classification error in the nearest neighbor search step of the algorithm.

5. IMAGE PREPROCESSING

To improve the performance of the LLE and NDA algorithms, we have analyzed the use of a preprocessing step in both algorithms. The NDA algorithm

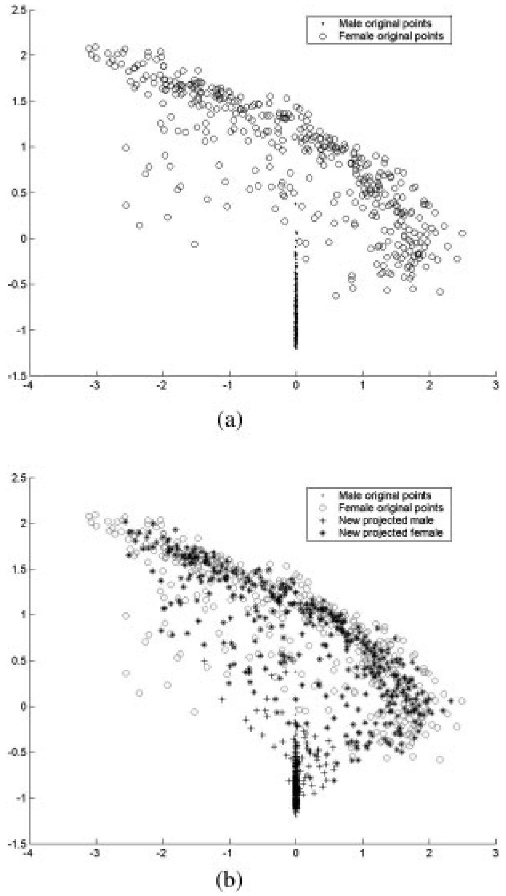


Figure 5. (a) Example of an embedding of two-class 500 data samples using supervised LLE (to a two-dimensional space). (b) The projection of 500 new unseen data samples.

tries to maximize the mean distances to the nearest neighbors of the extraclass vectors while it minimizes the distances between vectors of the same class. This criterion is very useful to discriminate and classify samples, but it can be seriously affected by the noise present in the largest part of the natural images. To solve this problem we have added an initial PCA dimensionality reduction step, which filters the noise preserving the largest part of the data variance. Using the first 300 principal components with the face data set, 98% of the variance is preserved, and we have seen that any choice from 200 to 400 principal components yields similar results after NDA reduction and NN classification. We realized experimentally that this stage considerably improves the results of NDA, so we have merged the NDA projection matrix with the PCA basis to construct the final projection.

6. GENDER RECOGNITION EXPERIMENT

We are facing a gender recognition problem, so we have tested our modification of the NDA algorithm and the LLE nonlinear projection for this purpose. We have also compared the results with other well-known classification algorithms.

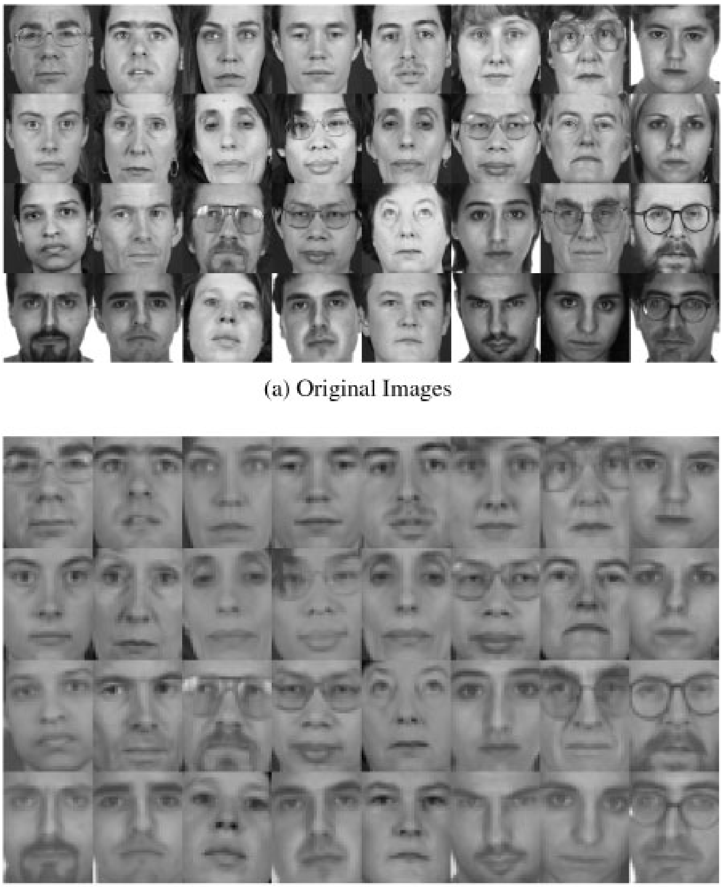
In this experiment, we used a face database composed of samples of two different internet-available faces databases, the AR Face database[25] and XM2VTS database.[26] As a preprocessing step, all images were aligned, manually selecting the center pixel of each eye and translating and resizing the face image with respect to eye distance. Once face images had both eyes in the same position, they were cropped to a 40-by-32 thumbnail. Then a global mean-variance normalization was performed. The final data set consisted of 3461 1280-dimensional vectors. As can be seen in Figure 6 we tried to avoid the presence of hair information in the final face database, which makes the problem more difficult to solve.[2]

The error rates shown in the results were estimated with fivefold cross validation. The 3461 data samples were divided into five sets, four of them used for training and the other one used for testing (we repeated it for the five sets and averaged the results). We have built the sets in a pseudorandom way, following three basic restrictions:

(1) The same number of males and females should appear in each set.
(2) Faces from the same person cannot be present in more than one set, to avoid person recognition instead of gender recognition.
(3) The number of faces of each database should be very similar for each group.

As can be seen in Table III, our modification of the NDA algorithm improves the results of classic dimensionality reduction techniques, PCA and FLD. We can see how FLD performs worse than nearest neighbor classification in the original space. This happens due to the limitations of FLD when non-Gaussian data are used, and due to the fact of dealing with a two-class problem, the resulting one-dimensional projected space is not enough to separate both classes. The NDA algorithm overcomes these drawbacks.

We have also tested a bagging procedure to improve the NDA results. Bagging is a specific technique to improve the performance of a classifier by combining several instances of the classifier. Two different techniques can be useful for this task, boosting[27] and bagging.[28] The goal of boosting is to combine a set of weak classifiers to get a classifier with better performance. On the other hand, bagging tries to improve the performance by bootstrapping the data samples in different sets and combining the results of the classification using each set, with some rule such as majority voting or simply averaging the results. The goal of bagging is to avoid or reduce the influence of misleading examples, because they can be isolated in a few sets, having low influence in the final voting results. In our classification scheme it is difficult to use boosting, because the hypothesis of a weak classifier assumed in boosting is not fulfilled. Bagging, on the other hand, can be very useful in the NDA scheme (as we will show in the results). We have broken the training set into subsets, and we have learned the NDA algorithm in



(a) Original Images

(b) Processed Images

Figure 6. Examples of face images used in the experiment before and after applying the mean-variance normalization.

Table III. Gender recognition accuracies.

Algorithm	Accuracy
NN	86.28
PCA	86.57
SVM	90.95
FLD	81.30
LLE	76.27
SLLE	87.12
NDA	90.56
Bagged NDA	91.76

each one using low dimensionality (≈10). Then we have used the NN rule to classify each projected sample, and we have combined the results using majority voting. The final result shows that bagging increases the performance of our scheme a little bit. The use of bagging also has another important advantage, the lower computational cost due to performing the dimensionality reduction using reduced subsets and to lower dimensionality spaces.

The training set has been divided into 49 subsets and we have trained an NDA reduction to a 16-dimensional space in each set. The final label assignation is obtained by majority voting. This technique has achieved the best results in our experiments, even slightly better than SVM. Another important advantage of the NDA algorithm is that it is able to reach the best performance even in very reduced spaces. As we can see in Figure 7, in a 20-dimensional space it reaches accuracies close to 90%, with low computational cost.

On the other hand, we can see how LLE is a bad algorithm to use in gender classification[1] unless we use the supervised version of the technique. Initially we thought that the preservation of the local configuration of each point would improve the results of a nearest neighbor classifier, but the results are far from the other analyzed techniques. As can be seen in Figure 3, LLE projection captures data related to global illumination, ethnicity, gesture, and beard, but it does not allow good gender discrimination with a nearest neighbor approach. We can also see in the results how the supervised version of LLE achieves performances close to the other analyzed techniques, especially in low dimensional subspaces.

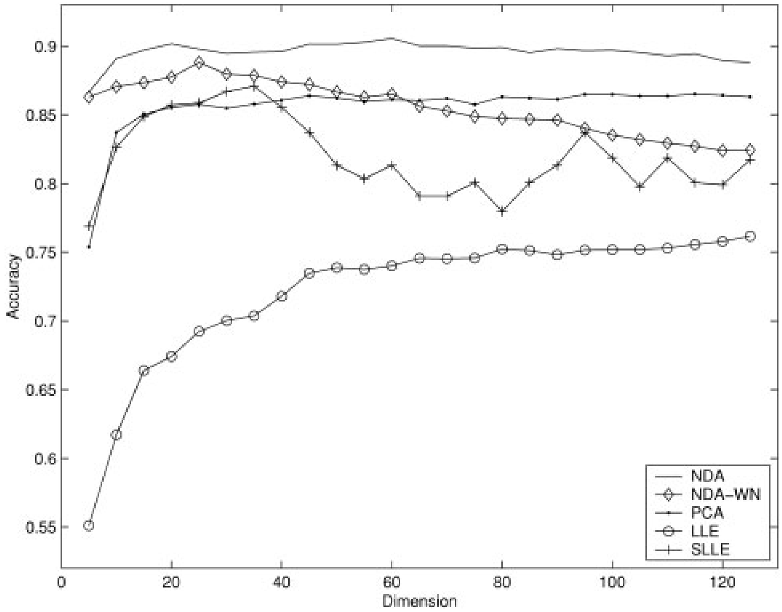


Figure 7. Recognition rate as a function of final dimension.

7. CONCLUSIONS

In this article we have compared different dimension reduction techniques especially suited for the nearest neighbor classification approach when used in gender recognition. In particular we have analyzed a modification of a nonparametric discriminant analysis algorithm, which obtains the best accuracies in our experiments. The results obtained show that the use of NDA allows the best classification rates even in low dimensional subspaces, which makes the algorithm computationally efficient.

Another important consideration is the use of bagging to improve the results, and even make the learning algorithm computationally more efficient due the fact of working with reduced subsets of the training data.

We also show the performance of the LLE algorithm for the same purpose. Initially, it could be thought that LLE could be the ideal representation for a nearest neighbor approach due to its property of conserving the local geometry of neighbor points. The results have shown that this technique achieves poor accuracies unless we use its supervised version.

Acknowledgments

This work is supported by grant TIC2003-00654 and FP2000-4960 Ministerio de Ciencia y Tecnologia, Spain.

References

1. Graf AB, Wichmann FA. Gender classification of human faces. Lect. Notes Comput Sci 2002;2525:491–501.
2. Moghaddam B, Yang MH. Learning gender with support faces. IEEE Trans Pattern Anal Mach Intell 2002;24:707–711.
3. Turk M, Pentland A. Eigenfaces for recognition. J Cogn Neurosci 1991;3:71–86.
4. Kirby M, Sirovich L. Application of the Karhunen-Loeve procedure for the characterization of human faces. IEEE Trans Pattern Anal Mach Intell 1990;12:103–108.
5. Hyvarinen A, Karhunen J, Oja E. Independent component analysis. New York: John Wiley and Sons; 2001.
6. Lee DD, Seung HS. Learning the parts of objects with nonnegative matrix factorization. Nature 1999;401:788–791.
7. Fisher R. The use of multiple measurements in taxonomic problems. Ann Eugenics 1936; 7:179–188.
8. Fukunaga K, Mantock J. Nonparametric discriminant analysis. IEEE Trans Pattern Anal Mach Intell 1983;5:671–678.
9. Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science 2000;290:2323–2326.
10. Okun O, Kouropteva O, Pietikäinen M. Supervised locally linear embedding algorithm. In: Proc 10th Finnish Artificial Intelligence Conference, Oulu, Finland, December 2001. pp 50–61.
11. Burton AM, Bruce V, Dench N. What's the difference between men and women? Evidence from facial measurement. Perception 1993;22:153–176.
12. Brunelli R, Poggio T. HyperBF networks for gender classification. In: Proc DARPA Image Understanding Workshop; 1992. pp 311–314.

13. Bruce V, Burton AM, Hanna E, Healey P, Mason O, Coombes A, Fright R, Linney A. Sex discrimination: How do we tell the difference between male and female faces. Perception 1993;22:131–152.
14. Kemp R, Pike G, White P, Musselman A. Perception and recognition of normal and negative faces—The role of shape from shading and pigmentation cues. Perception 1996;25: 37–52.
15. Abdi H, Valentin D, Edelman B, O'Toole A. More about the difference between men and women: Evidence from linear neural networks and the principal component approach. Perception 1995;24:539–562.
16. Cottrell G. Empath: Face, emotion, and gender recognition using holons. Adv Neural Inform Process Syst 1991;3:564–571.
17. Golomb BA, Lawrence DT, Sejnowski TJ. Sexnet: A neural network identifies sex from human faces. Adv Neural Inform Process Syst 1991;3:572–577.
18. Tamura S, Kawai H, Mitsumoto H. Male/female identification from 8 × 6 very low resolution face images by neural network. Pattern Recogn 1996;29:331–335.
19. Gutta S, Wechsler H, Phillips PJ. Gender and ethnic classification. In: Proc IEEE Int Conf on Automatic Face and Gesture Recognition; 1998. pp 194–199.
20. Devijver P, Kittler J. Pattern recognition: A statistical approach. London, UK: Prentice Hall; 1982.
21. Fukunaga K. Introduction to Statistical Pattern Recognition, 2nd ed. Boston, MA: Academic Press; 1990.
22. Saul LK, Roweis ST. Think globally, fit locally: Unsupervised learning of nonlinear manifolds. Technical report CIS-02-18. Philadelphia, PA: University of Pennsylvania; 2002.
23. Masip D, Vitria J. An experimental comparison of dimensionality reduction for face verification methods. Lect Notes Comput Sci 2003;2652:530–537.
24. de Ridder D, Duin RP. Locally linear embedding for classification. Technical report. Delft, the Netherlands: Delft University of Technology; 2002.
25. Martinez A, Benavente R. The AR face database. Technical Report 24. Computer Vision Center; 1998.
26. Matas J, Hamouz M, Jonsson K, Kittler J, Li Y, Kotropoulos C, Tefas A, Pitas I, Tan T, Yan H, Smeraldi F, Bigun J, Capdevielle N, Gerstner W, Ben-Yacoub S, Abdeljaoued Y, Mayoraz E. Comparison of face verification results on the XM2VTS database. In: Proc Int Conf on Pattern Recognition, July 1999. pp 4858–4864.
27. Schapire RE. A brief introduction to boosting. In: Proc Int Joint Conf on Artificial Intelligence, 1999. pp 1401–1406.
28. Skurichina M, Duin RPW. Bagging, boosting and the random subspace method for linear classifiers. Pattern Anal Appl 2002;5:121–135.