Revisiting Linear Discriminant Techniques in Gender Recognition

Juan Bekios-Calfa, José M. Buenaposada, and Luis Baumela

Abstract—Emerging applications of computer vision and pattern recognition in mobile devices and networked computing require the development of resource-limited algorithms. Linear classification techniques have an important role to play in this context, given their simplicity and low computational requirements. The paper reviews the state-of-the-art in gender classification, giving special attention to linear techniques and their relations. It discusses why linear techniques are not achieving competitive results and shows how to obtain state-of-the-art performances. Our work confirms previous results reporting very close classification accuracies for Support Vector Machines (SVMs) and boosting algorithms on single-database experiments. We have proven that Linear Discriminant Analysis on a linearly selected set of features also achieves similar accuracies. We perform cross-database experiments and prove that single database experiments were optimistically biased. If enough training data and computational resources are available, SVM's gender classifiers are superior to the rest. When computational resources are scarce but there is enough data, boosting or linear approaches are adequate. Finally, if training data and computational resources are very scarce, then the linear approach is the best choice.

Index Terms—Computer vision, gender classification, Fisher linear discriminant analysis.

1 Introduction

Demographic classification, and in particular gender recognition, is a research topic with a high application potential in areas such as surveillance, face recognition, video indexing, and dynamic marketing surveys. It has attracted the interest of researchers in computer vision and pattern recognition for years [1], [2], [3], [4], [5], [6], [7], SEXNET [5] being the first attempt to recognize gender from faces. Solutions to this problem may be broadly grouped into *appearance-based* approaches, and *feature-based* approaches. Appearance-based approaches use the cropped, resized, and illumination normalized texture of the whole face as a classification attribute. On the other hand, feature-based approaches are based on extracting a set of discriminative face features.

Moghaddam and Yang [1] introduced the best gender recognition algorithm in terms of reported classification rate. They adopted an appearance-based approach with a classifier based on a Support Vector Machine with Radial Basis Function kernel (SVM+RBF) [1]. They reported a 96.6 percent recognition rate for classifying 1,775 images from the FERET database using automatically aligned and cropped images and a fivefold cross-validation. Baluja and Rowley [2] report a bias in the previous estimation caused by the existence of subjects with the same identity in different folds. In the same experiment they achieved a 93.5 percent success rate using SVM+RBF with manual alignment and a proper cross-validation.

Feature-based approaches use pixel-wise gray-level differences [2], Haar-like wavelets [3], [6], multiscale filter banks [7], or Locally Binary Patterns (LBP) [3], [4] to recognize the gender of a face. Shakhmarovich et al. [6] achieved 79 percent and 79.2 percent recognition accuracy in gender and ethnicity classification, respectively, on a set of difficult images obtained from the Web. They used Haar-like features within an AdaBoost-based approach, which is several orders of magnitude faster than SVM. Baluja and Rowley [2] used pixel-wise gray-level comparisons as weak classifiers within an AdaBoost learning scheme. They used manually aligned images from the Color FERET database "fa" and "fb" galleries and achieved 94 percent recognition accuracy. Their classifier is approximately 50 times faster than Moghaddan and Yang's SVM solution [1].

Recently, Mäkinen and Raisamo [3] performed a set of experiments using 411 images (304 for training and 107 for testing) from the FERET database. They compared appearance-based, feature-based, aligned, and unaligned approaches, among others. They got similar performance results for feature-based AdaBoost and appearance-based SVM+RBF classifiers. In another work [4] they experimented with different databases, classifier combination, and face normalizations.

With the notable exception of [2], existing approaches to gender recognition focus mainly on high-performance computer systems. Emerging applications of video analysis in mobile devices and networked computing have recently attracted interest in the development of computer vision and pattern recognition algorithms for resource-limited devices. Linear classification techniques have an important role to play given their simplicity and low computational requirements at runtime. In this paper, we revisit and compare various linear classification algorithms. We prove that, with a linear feature selection, these approaches achieve results comparable to the best gender classifiers based on SVM+RBF [1] and Boosting [2]. Moreover, in the context of very limited data and computational resources, they achieve the best generalization.

2 Linear Discriminant Analysis (LDA)

Given a multiclass classification problem with c classes and p sample points, $\{\mathbf{x}_i\}_{i=1}^{p}$, LDA provides a linear projection of the initial samples onto a subspace of at most $d = c - 1$ dimensions, maximizing the ratio of the between-class and within-class separation. The basis of the transformed subspace, $\{\mathbf{w}_i\}_{i=1}^{d}$, is obtained by maximizing $J(w) = \sum_{i=1}^{d} \frac{\mathbf{w}_i^\top \mathtt{S}_B \mathbf{w}_i}{\mathbf{w}_i^\top \mathtt{S}_W \mathbf{w}_i}$, where $\mathtt{S}_B$ and $\mathtt{S}_W$ are, respectively, the between-class and within-class scatter matrices [8]. The maximum is given by the following generalized eigenvalue equation: $\mathtt{S}_B\mathtt{W} = \mathtt{S}_W\mathtt{W}\mathtt{D}$, where $\mathtt{W}$ is a matrix whose columns are $\mathtt{w}_i$ and $\mathtt{D}$ is the diagonal matrix of eigenvalues. The rank of matrix $\mathtt{S}_B$ is at most $c - 1$ and, generally, so is the rank of the LDA projection matrix, $\mathtt{W}$.

In the following sections, we describe three linear dimensionality reduction techniques directly related to LDA, which we have compared in our experiments.

2.1 LDA in the PCA-Transformed Subspace (I), PCA+LDA

When dealing with image classification problems, it is very common to have fewer sample vectors (images) than features (pixels). In such cases, the within-class scatter matrix $\mathtt{S}_W$ is singular and the LDA projection matrix $\mathtt{W}$ cannot be computed. Since the covariance matrix of the full sample set is $\mathtt{S}_m = \mathtt{S}_B + \mathtt{S}_W$, an alternative solution is given by using $\mathtt{S}_m$ instead of $\mathtt{S}_W$ in the previous eigenvalue equation [8]. In this case, performing Principal

- *J. Bekios-Calfa is with the Departamento de Ingeniería de Sistemas y Computación, Universidad Católica del Norte, Avenida Angamos 0610, Gran Vía, Antofagasta, Chile. E-mail: juan.bekios@ucn.cl.*
- *J.M. Buenaposada is with the Departamento de Ciencias de la Computación, Universidad Rey Juan Carlos, C/ Tulipan s/n, 28933 Móstoles, Spain. E-mail: josemiguel.buenaposada@urjc.es.*
- *L. Baumela is with the Departamento de Inteligencia Artificial, Universidad Politécnica de Madrid, Campus de Montegancedo s/n, 28660 Boadilla del Monte, Spain. E-mail: lbaumela@fi.upm.es.*

Manuscript received 13 Feb. 2010; revised 13 Aug. 2010; accepted 22 Oct. 2010; published online 24 Nov. 2010.
Recommended for acceptance by M.-H. Yang.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2010-02-0097.
Digital Object Identifier no. 10.1109/TPAMI.2010.208.

Component Analysis (PCA) retaining only the eigenvectors associated with nonzero eigenvalues and then performing LDA in the transformed PCA subspace is equivalent to performing LDA in the original subspace [9]. Thus, when eigenvectors associated with nonzero eigenvalues are discarded, PCA+LDA will not be strictly equivalent to the global LDA counterpart. From now on, we will call this method PCA+LDA, irrespective of the eigenvectors chosen from the PCA step.

In most of our cases, we will have more training samples than pixels in the sample images. Therefore, all of the eigenvalues from PCA will be nonzero. Depending on the amount of training data, the classifier performance decreases when retaining all eigenvectors associated with nonzero eigenvalues (see Figs. 5 and 6). Thus, a crucial step here is to choose which PCA eigenvectors to keep so that no discriminant information is lost.

We select the dimension of the subspace resulting from the PCA step using a cross-validation scheme instead of the usual approach based on retaining the eigenvectors accounting for a given percentage of the variance (usually 95 percent or 99 percent) [11]. We sort PCA eigenvectors in descending eigenvalue order. We then perform cross-validation and select the dimension with the best performance. In Algorithm 1, N is the number of pixels in an image, $\mathtt{X}$ is a matrix with one sample image per column, $\mathbf{L}$ is the vector with the corresponding class labels (male or female), $\mathtt{P}_{PCA}$ is the PCA basis matrix sorted in decreasing order of eigenvalues, $\mathbf{I}$ is the PCA mean, R is the best classification rate obtained, and K is the best dimension.

Algorithm 1. PCA+LDA training

Input: $\mathtt{X}, \mathbf{L}$

Result: R, K

1: $R \Leftarrow 0$ {Initialize best classification rate to 0}
2: $K \Leftarrow 1$ {Initialize best feature number to 1}
3: Divide X in l folds $F = \{\mathtt{X}_1, \ldots, \mathtt{X}_l\}$.
4: **for** $j = 1$ to N **do** {l-fold cross-validation with j features retained before LDA}
5: **for** $i = 1$ to l **do**
6: $\mathtt{X}_{test} \Leftarrow \mathtt{X}_i$ {Test with fold i}
7: $\mathtt{X}_{train} \Leftarrow F - \{\mathtt{X}_i\}$ {Train with the rest of folds}
8: $[\mathtt{P}_{PCA}, \mathbf{I}] \Leftarrow$ PCA($\mathtt{X}_{train}$) {Principal Component Analysis}
9: $\mathtt{B}$ is assigned the first j columns in $\mathtt{P}_{PCA}$.
10: $\mathtt{Y} \Leftarrow \mathtt{B}^\top(\mathtt{X}_{train} - [\mathbf{I} \ldots \mathbf{I}])$ {Projection onto PCA subspace}
11: $\mathtt{P}_{LDA} \Leftarrow LDA(\mathtt{Y})$ {Fisher Linear Discriminant Analysis}
12: $\mathtt{Z} \Leftarrow \mathtt{P}_{LDA}\mathtt{Y}$ {Projection onto LDA subspace}
13: C $\Leftarrow$ trainBayesianClassifier($\mathtt{Z}, \mathbf{L}$)
14: $\mathtt{Z}_{test} \Leftarrow \mathtt{P}_{LDA}\mathtt{B}^\top(\mathtt{X}_{test} - [\mathbf{I} \ldots \mathbf{I}])$
15: $r_i \Leftarrow$ classify(C, $\mathtt{Z}_{test}, \mathbf{L}$)
16: **end for**
17: $R_j \Leftarrow \frac{1}{l}\sum_{i=1}^{l} r_i$
18: **if** $R_j > R$ **then**
19: $R \Leftarrow R_j$
20: $K \Leftarrow j$
21: **end if**
22: **end for**

We will show in the experiments that this feature selection process is essential to getting state-of-the-art performance with the PCA+LDA procedure. This is not the first time that this kind of approach has been used in the literature. In their comparison of PCA and LDA approaches for appearance-based object recognition, Martínez and Kak also select the best PCA dimension prior to performing LDA [12].

2.2 LDA in the PCA-Transformed Subspace (II), PCA-M+LDA

An alternative way of selecting the PCA eigenvectors is to sort them according to their agreement with matrix $\mathtt{S}_B$ [10]. In this case, we give more importance to the eigenvectors that are parallel to the subspace spanned by the class means. The importance of an $\mathtt{S}_m$ eigenvector, $\mathbf{u}_j$, is then given by $I_j = \sum_{i=1}^{q}(\mathbf{u}_j^\top \mathbf{v}_i)^2$, $q = rank(\mathtt{S}_B)$, where $\mathbf{v}_i$ are the eigenvectors of $\mathtt{S}_B$ [10].

With PCA-M+LDA we denote the algorithm that performs PCA, then sorts the PCA eigenvectors by decreasing value of I_j, chooses the first k eigenvectors in the new order, and finally performs LDA. In the PCA-M+LDA case, our training procedure is as shown in Algorithm 1 but with an important difference: After PCA (line 8) in PCA-M+LDA, we sort $\mathtt{P}_{PCA}$ columns by decreasing value of I_j.

2.3 LDA in the ICA-Transformed Space (ICA+LDA)

ICA tries to explain the original sample data in terms of statistically independent random vectors. Let $\mathtt{X}$ be a data matrix whose columns are the sample vectors. Linear ICA algorithms find a matrix $\mathtt{P}$ that projects $\mathtt{X}$ onto an independent components subspace, $\mathtt{S} = \mathtt{P}\mathtt{X}$.

Most researchers using ICA-based results use either FastICA or Infomax algorithms [13]. These procedures search for vectors $\mathbf{v}_i$, rows of matrix $\mathtt{P}$ such that the rows of $\mathtt{S}$ have maximally non-Gaussian distribution and are mutually (approximately) uncorrelated. A simple way to achieve this objective is to make PCA, retain only eigenvectors with nonzero eigenvalues, whiten, and then search for a rotation matrix $\mathtt{R}, \mathtt{S} = \mathtt{R}^\top \Lambda^{-1}\mathtt{B}^\top\mathtt{X} = \mathtt{R}^\top\mathtt{Z}$, where $\mathtt{Z}$ are the whitened PCA projections of sample vectors in $\mathtt{X}$ and $\mathtt{B}$ are the eigenvectors associated with the Λ diagonal matrix with nonzero eigenvalues [13].

Abusing the concept of independence, some approximations use the independent components obtained by ICA as a basis for expanding a linear subspace [14]. We have used Algorithm 1 to train the FastICA+LDA classifier. The only difference is that now lines 8 to 10 use FastICA to estimate projection matrix $\mathtt{P}$, selecting the first j rows of $\mathtt{P}$ and projecting $\mathtt{X}_{train}$ onto the first j FastICA features to obtain matrix $\mathtt{Y}$.

3 Experiments

In this section, we evaluate the performance and compare the above linear approaches with the best nonlinear classifier, SVM+RBF, as used by Moghaddam and Yang [1], and Baluja and Rowley's [2] pixel-wise boosting-based algorithm.

We have used one nonpublic database from the Universidad Católica del Norte (UCN) in Chile, termed the UCN database, the Color FERET database [15], the Productive Aging Lab Face (PAL) database from the University of Texas at Dallas [16], and the same training and test images sets used by Mäkinen and Raisamo [3] from the Gray FERET database:

- The UCN database is a nonpublic database consisting of mug-shot images (one per individual) of students and academics from UCN. They have been acquired with different imaging devices under different resolutions, illumination conditions, and in which faces are not strictly frontal (see Fig. 1). There are 10,700 individuals, 5,646 male and 5,054 female. In our experiments, we use 5,628 male and 5,041 female images since the face detector missed some faces when preparing the data.
- The Color FERET Face database is a publicly available resource for face analysis research. It has multiple images of 994 individuals, 591 male and 403 female. For our experiments, we use one image per subject from the FERET database *fa* gallery. From this gallery we only employ

Fig. 1. Some cropped and resized images, after face detection, from the UCN database.

Fig. 2. Some cropped and resized images, after face detection, from the Color FERET database, *fa* gallery.

Fig. 3. Some cropped and resized images, after face detection, from the PAL database.

Fig. 4. The first row displays raw cropped face images using the face detector. The second shows equalized and masked images.

402 female images since the face detector missed one female face (see Fig. 2).

- The PAL Database consists of frontal pictures of 576 individuals. The right profile and some facial expressions are also available for some subjects. There are 219 male and 357 female subjects divided into four groups depending on their age: 18-29, 30-49, 50-69, and 70-93 (see Fig. 3).
- The subset of the Gray FERET database used by Mäkinen and Raisamo [3] in their out-of-plane face rotation sensitivity analysis experiment[1]consists of 304 frontal images for training and 1,008 for testing organized in 9 different orientations from +60 to -60 degrees.

Before classifying we crop and resize images to 25×25 pixels using OpenCV's[2] 2.0.0 face detector, which is based on [17]. For manual alignment we use the location of the eyes and mouth center. Additionally, we perform a histogram equalization in order to gain some independence from illumination changes. Finally, we also apply an oval mask to prevent the background from influencing our results (see Fig. 4).

In all LDA-based experiments, we use a simple Bayesian classifier assuming Gaussian distribution (see Fig. 10 to verify that this is a reasonable assumption). For the SVM+RBF tests, we train the classifier using the Sequential Minimal Optimization [18] algorithm implemented in WEKA Explorer.[3] We search for the best C (trade-off between margin and training error) and gamma (RBF radius) parameters in a grid of different combinations. In all databases, except for UCN, C varies from 1 to 991 in steps of 10 (99 samples) and gamma varies from 0.001 to 0.1 in steps of 0.001 (100 samples). For UCN, C took values in {1, 10, 100, 1,000} and gamma in {0.001, 0.002, 0.004, 0.006, 0.008}. In Table 1 we provide the best parameters for each database. We implemented the pixel-wise boosting algorithm of Baluja and Rowley as described in their paper, using 1,000 weak classifiers chosen evaluating 1 percent of all possible weak classifiers [2]. We called it Baluja1000. Additionally, for comparison purposes, we also trained a Baluja625 classifier, which matches the computational complexity of the linear classifiers.

3.1 Single Database Tests

In the first experiment, we perform single-database tests using a five-fold cross-validation scheme. In Table 1, we show the results of this experiment. All face images are unaligned except for the FERET database, for which we perform both unaligned (FERET column) and manually aligned (FERET Align column) tests. Subjects with the same identity are kept in one fold since we use only one image per subject.

In the following paragraphs, we discuss these experimental results:

- **Manually aligned versus unaligned faces.** One first obvious result from Table 1 is that we have not found a significant performance difference between manually aligning the images or just classifying them after detection. This confirms similar previous results in [3]. This is due to precision achieved by the face detector. On average, it achieves an accuracy of about half a pixel for frontal 25×25 resized faces from FERET fa gallery.
- **LDA classification.** LDA achieves in FERET a 77.68 percent success rate. We did not test LDA on PAL because $\mathtt{S}_m$ is rank deficient, given the small number of samples. This experiment confirms the poor results obtained by Moghaddam and Yang [1].

  On the other hand, LDA on the UCN database, with around 10,000 images, provides a success rate of 92.65 percent. We can conclude that, because of the curse of dimensionality, 993 images do not provide enough information for LDA to find the right projection from a 625 dimensional space. Increasing the number of training images to 10,000 (UCN database) provides enough data for LDA to become a competitive classifier.

  So, when the dimensionality of the problem is high (625 dimensions in our case) compared to the number of samples (994), LDA does not provide a good solution, even if $\mathtt{S}_m$ is a full rank matrix.
- **PCA+LDA classification.** We have tested both PCA+LDA and PCA-M+LDA using Algorithm 1 to select the dimension of the PCA subspace. In Table 1 we report the results achieved by this iterative procedure. With a linear feature selection before LDA, PCA+LDA and PCA-M+LDA achieve a performance competitive with the state-of-the-art. PCA-M+LDA and PCA+LDA plots in Fig. 5 confirm that when the size of the database is large (UCN) the performance of the classifier does not depend so much on the dimension of the intermediate PCA subspace. In this case, we can safely select all dimensions of this subspace. This is equivalent to performing LDA on the original data.

  The reason for the PCA-M+LDA and PCA+LDA algorithms being so successful is that they diminish LDA's *curse of dimensionality* by selecting only PCA's most discriminant directions. In our problem, PCA-M+LDA

1. Available online at http://www.cs.uta.fi/hci/mmig/vision/datasets/.
2. http://opencv.willowgarage.com.
3. http://www.cs.waikato.ac.nz/ml/weka.

TABLE 1
Classification Rates and Standard Deviation for Single Database Five-Fold Cross-Validation Tests

Classifier	Data Base			
	FERET	FERET Align	PAL	UCN
SVM+RBF	93.95±2.60% (247) C=100, $\gamma = 0.001$	93.46±1.65% (314) C=10, $\gamma = 0.003$	89.81±1.55% (320) C=20, $\gamma = 0.01$	95.39±0.21% (1891) C=100, $\gamma = 0.002$
PCA+LDA	93.33±2.33% (130)	93.57±1.39% (120)	85.52±3.01% (180)	92.86±0.64% (460)
PCA-M+LDA	92.83±0.75% (100)	93.57±1.25% (60)	84.83±1.98% (140)	92.86±0.77% (300)
ICA+LDA	93.33±2.33% (130)	93.57±1.39% (120)	85.52±3.01% (180)	92.86±0.64% (460)
LDA	77.68±2.61%	77.09±2.26%	——	92.65±0.65%
Baluja625	92.12±1.36%	93.17±1.65%	85.86±2.48%	93.87±0.64%
Baluja1000	93.33±1.06%	93.07±1.99%	87.24±1.27%	94.67±0.30%

In the SVM and LDA rows, respectively, we show in parentheses the number of support vectors and the number of features kept before performing LDA. For SVM we also show the best C and γ parameters.

and PCA+LDA perform equally well. This is because the most discriminant eigenvectors are the ones with the largest variance. This makes sense since we are dealing with constant lighting, neutral expression, and frontal face images. In this case, most variability comes from person-to-person differences and thus gender appearance variation explains most of the variance in the data. Note that, in general, this might not be true.

In conclusion, with an appropriate feature selection procedure LDA can achieve a competitive classification performance and overcome the curse of dimensionality.

- **ICA+LDA classification.** We now analyze the result of estimating the intermediate subspace using ICA instead of PCA. Again using Algorithm 1, we estimate the dimension of this subspace. In Table 1 and Fig. 6 we show the results of this experiment. Not surprisingly, the results for ICA+LDA and PCA+LDA are very similar. Moreover, the ICA+LDA and PCA+LDA approaches have the same classification rates when the number of selected features is large enough (see Fig. 7). These results were theoretically predictable. Since FastICA is equivalent to a whitened PCA plus a rotation [13] and LDA is a rotation invariant dimensionality reduction technique, then PCA+LDA and FastICA+LDA are equivalent when there is no feature selection.

Jain and Huang reported a 99.3 percent success rate in an experiment using FastICA+LDA and a euclidean classifier [14]. They tested their approach with 500 images from the FERET database. For training they used 200 images (100 male and 100 female). The remaining 300 images (150 male and 150 female) were used for testing. A possible reason for the discrepancy between their result and ours is the small size of the database used, which may have biased their evaluation.

The last issue considered in this set of tests is the sensitivity of classifiers to out-of-plane face rotation. We use the 304 frontal face images from Mäkinen and Raisamo's subset of the Gray FERET database for training. We test with the set of 1,008 images taken at different horizontal face orientations. For this experiment we manually align the faces since, for extreme angles, the face detector misses most of them. In Fig. 8, we show the results of this test. The asymmetry of the plot is caused by some training images being slightly rotated toward the negative angles. On average, all methods perform similarly, showing a higher performance on the negative rotation side. Boosting-based algorithms perform slightly better for the negative range, whereas linear methods are marginally ahead in the positive.

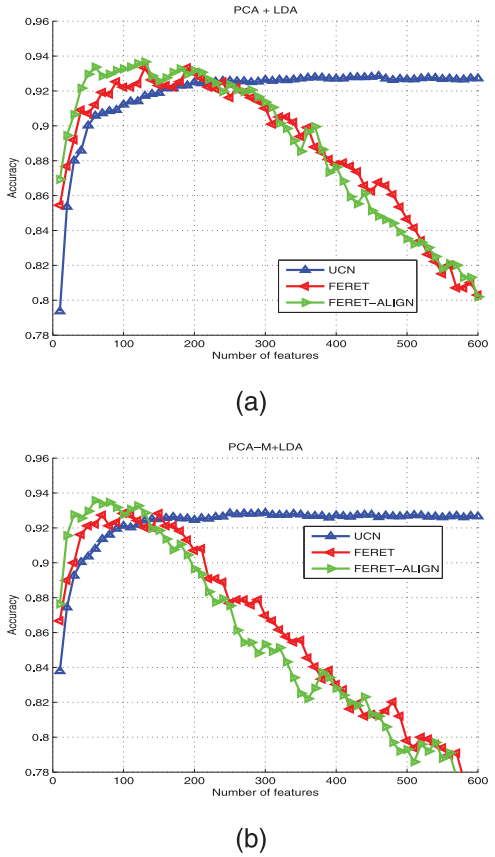


Fig. 5. Classification performance (variable R_j in Algorithm 1) as the dimension of the intermediate PCA subspace increases for (a) PCA+LDA and (b) PCA-M+LDA.

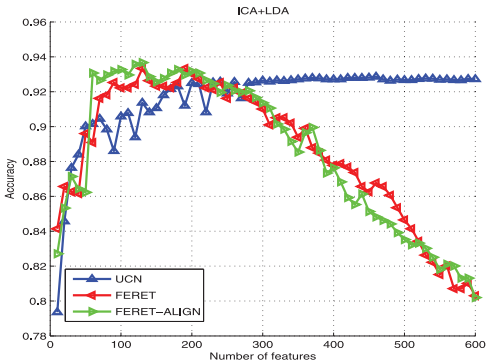
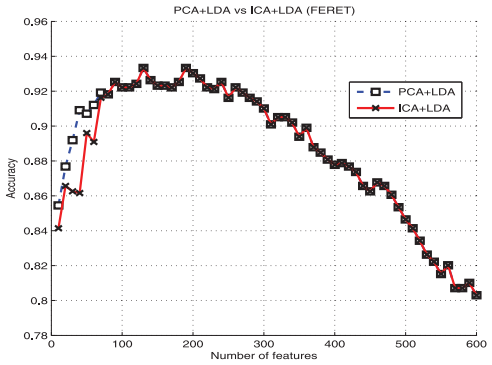
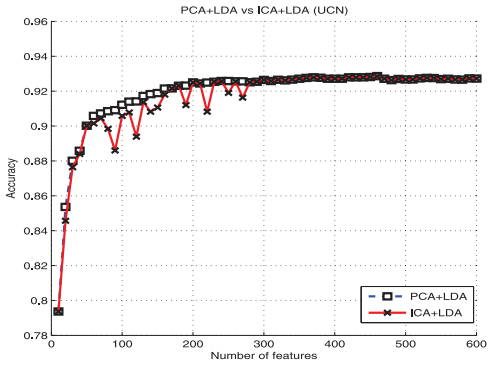


Fig. 6. Classification performance (variable R_j in Algorithm 1) as the dimension of the intermediate ICA subspace increases for ICA+LDA.



(a)



(b)

Fig. 7. Comparison of PCA+LDA and ICA+LDA for (a) the FERET and (b) the UCN databases.

On average, the performance of all classifiers in single-database experiments is very similar. The most significant difference is achieved by SVM+RBF on the large UCN database.

3.2 Cross-Database Tests

In the second experiment, we perform cross-database classification tests. With this experiment we get an idea of the generalization capabilities of the classifiers. We use all of the images from one database for training and the images from another for testing. Here we use the classifier parameters obtained in the single database experiments. Table 2 shows the results.

FERET and UCN databases have similar demography but different acquisition conditions. Cross-database tests with FERET and UCN provide results analogous to the single-database tests discussed in Section 3.1, with an overall decrease in performance most noticeable when training with FERET and testing with UCN (*FERET/UCN*), caused by the more general acquisition conditions in UCN. Here all classifiers have close performances, SVM+RBF

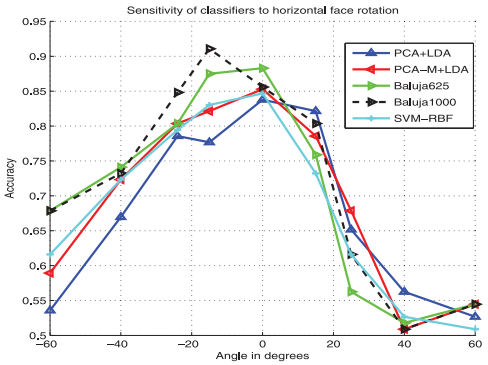


Fig. 8. Sensitivity analysis to out-of-plane face rotation.

and Baluja1000 being marginally better, respectively, when training with a large (*UCN/FERET*) and a small database (*FERET/UCN*). Also, *FERET/UCN* results are quite poor for the LDA classifier. However, when training with UCN and testing with FERET (*UCN/FERET*), the LDA approach can compete with the other classification procedures. This confirms our previous single-database experiment.

Cross-database tests involving PAL, FERET, and UCN are more demanding in terms of classifier generalization capabilities since the demography in PAL is quite different from that in the FERET and UCN databases. It includes people from more ethnic groups and a broader range of ages (see Figs. 1, 2 and 3). FERET and UCN are demographic subsets of PAL. If the training database is large, the SVM+RBF classifier clearly achieves the best performance (*UCN/PAL*). When training with small databases (PAL and FERET), performance differences become narrower. In the more challenging cases, which are *FERET/PAL*, given the narrow demography and controlled acquisition in FERET, and *PAL/UCN*, for the general acquisition conditions in UCN, SVM+RBF performance is slightly behind boosting and two-stage linear classifiers, the latter being marginally ahead in these cases. In the PAL/FERET test, SVMs and boosting approaches perform marginally better, in spite of PAL being a small database; this is possibly due to the broad demography in PAL.

These experiments, together with the sensitivity analysis in Section 3.1, seem to suggest that linear classifiers have the best generalization in situations where training data is very scarce and with low variability (e.g., narrow demography). To confirm this hypothesis we have performed one more experiment training with Mäkinen's FERET *pose-ba* gallery, which contains 112 frontal face images (56 male and 56 female), mostly Caucasian, and testing with PAL and UCN databases. We trained the classifiers using the same procedure as in Section 3.1. Classifier parameters are now, SVM+RBF ($C = 10$, $\gamma = 0.007$), PCA+LDA (37 features), PCA-M+LDA (80 features). In this case, see *Mak-ba/PAL* and

TABLE 2
Classification Rates for Cross-Database Experiment (Train DB/Test DB)

Classifier	Training/Testing							
	FERET/UCN	UCN/FERET	FERET/PAL	UCN/PAL	PAL/FERET	PAL/UCN	Mak-ba/PAL	Mak-ba/UCN
SVM+RBF	81.29%	91.03%	67.53%	79.27%	78.65%	74.09%	64.07%	60.86%
PCA+LDA	80.90%	88.72%	70.64%	72.88%	74.32%	76.53%	75.47%	72.11%
PCA-M+LDA	80.35%	88.92%	71.50%	73.23%	76.13%	76.09%	70.12%	67.16%
LDA	72.99%	88.72%	63.73%	73.75%	——	——	——	——
Baluja625	83.75%	90.72%	68.39%	74.61%	77.14%	75.18%	71.50%	65.24%
Baluja1000	84.18%	89.85%	70,12%	73.57%	78.85%	76.23%	65.53%	61.43%

Fig. 9. Some examples of wrongly classified images when training with FERET and testing with PAL. First row, male images classified as female. Second row, female images classified as male.

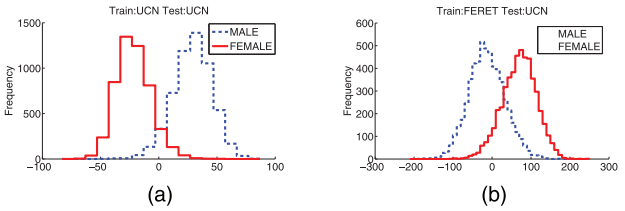


Fig. 10. Histograms showing the projection of the UCN database onto the PCA+LDA subspace computed with (a) the UCN and (b) the FERET databases.

Mak-ba/UCN columns in Table 2, performance differences among classifiers are larger. SVM+RBF has the lowest performance. PCA+LDA is clearly ahead of the rest.

We now analyze some classification errors in the *FERET/PAL* experiment (see Fig. 9). They are caused mainly by ages or ethnicities not present in the training data. For example, since FERET has few elderly female images, all elderly samples from PAL are classified as male (see the first two images in the second row of Fig. 9). Also, most images in FERET are from white Caucasians. Consequently, it is more likely that samples in PAL belonging to other ethnicities are wrongly classified (see Fig. 9). Finally, even with the ethnicities and ages present in the training data, there are faces that are difficult to classify (see the first two images in the first row or the third image in the second row of Fig. 9).

3.3 Computational Issues

Performance in terms of classifier success is not the only important issue in face analysis. Computational cost is also a key factor when processing millions of images [2] or when implementing these algorithms in a computing device such as a mobile phone or an IP camera. In terms of computational cost, Baluja et al.'s [2] pixel-wise boosting algorithm and the linear classifiers are the fastest gender recognition algorithms reported so far. Baluja's algorithm uses pixel-wise gray-level comparisons, a feature that is very fast to compute. For example, classifier Baluja625 would make on the order of 625 operations to classify one image. In the linear classifier case, the size of the projection matrix for the two-stage algorithms is independent of the intermediate PCA dimension. PCA+LDA and PCA-M+LDA projection matrices are, in fact, a single row vector with as many components as the number of image pixels. Classification is the result of thresholding the projected image, which results also in $1 \times 25 \times 25 = 625$ operations to classify one image.

Classification with an SVM+RBF classifier is a great deal more demanding in terms of computer operations. If we consider the smallest number of support vectors used by the SVM+RBF classifier in Table 1, 247 support vectors, SVM+RBF needs to make $247 \times 25 \times 25 = 154{,}375$ pixel operations to classify one image. For the UCN database the number of pixel operations is $1{,}891 \times 25 \times 25 = 1{,}181{,}875$, orders of magnitude larger than linear and boosting approaches.

4 Conclusions

In this paper we have reviewed the state-of-the-art in gender recognition. In single-database experiments, our work confirms previous results reporting similar classification accuracies for SVMs and pixel-wise boosting algorithms [2], [3], the former being slightly better when training with a large database. We have proven that linear techniques may also achieve similar accuracies in this context. We have experimentally confirmed that linear techniques based on ICA+LDA are equivalent to PCA-M+LDA and PCA+LDA. This is not surprising since most algorithms for ICA are equivalent to whitened PCA plus a rotation.

We have experimentally proven that single-database experiments are optimistically biased since the demography and acquisition conditions are usually very similar in the images of a database and they have an important impact in the performance of the classifier. Differences arise when classifiers are trained and tested with different databases. If there are 10,000 or more training samples, SVM+RBF is the best classifier. In the tough *UCN/PAL* test, it roughly achieves 80 percent success, at the expense of requiring 10^6 pixel operations to classify one image. If, on the other hand, we have time or computational constraints, boosting and linear approaches roughly achieve 75 percent success for this experiment with only 625 operations. If there are fewer training data (500 to 1,000 samples) with a broad demography, then all tested approaches achieve similar classification accuracies. Finally, if training data is scarce (300 images or less) and with a narrow demography, the PCA+LDA approach is the best choice. The success of simple linear techniques in this context is possibly caused by the high dimensionality of the input data space, which makes a kernelization step to achieve linear separability unnecessary.

We have found experimental evidence that supports the existence of a correlation between different demographic variables such as gender, age, and ethnicity. When a gender classifier is trained with a data set with limited demography (like the FERET or UCN databases) and then tested with a data set with more general samples (like the PAL database), the classification rate drops significantly. Dependencies between gender estimation and age [19] or ethnicity [20] have also recently been reported. New venues for research on gender in particular or demographic classification in general should take into account the relations between gender, age, and ethnicity variables in order to improve the classification across different age and ethnic groups.

Much research over recent years has focused on solving the linear discriminant analysis problem when S_m or S_W are singular matrices, e.g., [21], [22], [23]. From our experiments we can conclude that small sample size data can seriously compromise the performance of the linear discriminant classifier, even if the covariance matrices are not singular. We have experimentally proven that choosing the correct dimension for the intermediate subspace projection in a two-stage LDA algorithm improves the performance.

Acknowledgments

The authors gratefully acknowledge funding from the Spanish *Ministerio de Ciencia e Innovación* under contracts TIN2008-06815-C02-02 and TIN2010-19654 and the *Consolider Ingenio* program contract CSD2007-00018. Portions of the research in this paper use the FERET database of facial images collected under the FERET program, sponsored by the US Department of Defense (DOD) Counterdrug Technology Development Program Office.

References

[1] B. Moghaddam and M.-H. Yang, "Learning Gender with Support Faces," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 707-711, May 2002.

[2] S. Baluja and H.A. Rowley, "Boosting Sex Identification Performance," *Int'l J. Computer Vision,* vol. 71, no. 1, pp. 111-119, Jan. 2007.
[3] E. Mäkinen and R. Raisamo, "Evaluation of Gender Classification Methods with Automatically Detected and Aligned Faces," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 30, no. 3, pp. 541-547, Mar. 2008.
[4] E. Mäkinen and R. Raisamo, "An Experimental Comparison of Gender Classification Methods," *Pattern Recognition Letters,* vol. 29, no. 10, pp. 1544-1556, July 2008.
[5] B.A. Golomb, D.T. Lawrence, and T.J. Sejnowski, "Sexnet: A Neural Network Identifies Sex from Human Faces," *Advances in Neural Information Processing Systems,* pp. 572-577, Morgan Kauffmann, 1990.
[6] G. Shakhnarovich, P.A. Viola, and B. Moghaddam, "A Unified Learning Framework for Real Time Face Detection and Classification," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 16-26, 2002.
[7] A. Lapedriza, M.J. Marin-Jiménez, and J. Vitrià, "Gender Recognition in Non Controlled Environments," *Proc. Int'l Conf. Robotics and Automation,* pp. 834-837, 2006.
[8] K. Fukunaga, *Introduction to Statistical Pattern Recognition.* Academic Press, 1990.
[9] J. Yang and J.-y. Yang, "Why Can LDA Be Performed in PCA Transformed Space?" *Pattern Recognition,* vol. 36, pp. 563-566, 2003.
[10] M. Zhu and A.M. Martínez, "Selecting Principal Components in a Two-Stage LDA Algorithm," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. I, pp. 132-137, 2006.
[11] R. Johnson and D. Wichern, *Applied Multivariate Statistical Analysis.* Prentice-Hall, 1998.
[12] A.M. Martinez and A.C. Kak, "PCA versus LDA," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 23, no. 2, pp. 228-233, Feb. 2001.
[13] M.A. Vicente, P.O. Hoyer, and A. Hyvärinen, "Equivalence of Some Common Linear Feature Extraction Techniques for Appearance-Based Object Recognition Tasks," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 29, no. 5, pp. 896-900, May 2007.
[14] A. Jain and J. Huang, "Integrating Independent Components and Linear Discriminant Analysis for Gender Classification," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 159-163, 2004.
[15] P. Phillips, H. Moon, P. Rauss, and S. Rizvi, "The Feret Evaluation Methodology for Face Recognition Algorithms," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 10, pp. 1090-1104, Oct. 2000.
[16] M. Minear and D.C. Park, "A Lifespan Database of Adult Facial Stimuli," *Behavior Research Methods, Instruments and Computers,* vol. 36, pp. 630-633, 2004.
[17] P. Viola and M.J. Jones, "Robust Real-Time Face Detection," *Int'l J. Computer Vision,* vol. 57, no. 2, pp. 137-154, May 2004.
[18] J.C. Platt, "Fast Training of Support Vector Machines Using Sequential Minimal Optimization," *Advances in Kernel Methods: Support Vector Learning,* pp. 185-208, MIT Press, 1999.
[19] G. Guo, C.R. Dyer, Y. Fu, and T.S. Huang, "Is Gender Recognition Affected by Age?" *Proc. IEEE Int'l Conf. Computer Vision Workshop Human-Computer Interaction,* pp. 2032-2039, 2009.
[20] H. Ai and G. Wei, "Face Gender Classification on Consumer Images in a Multiethnic Environment," *Proc. Conf. Advances in Biometrics,* 2009.
[21] P. Zhang, J. Peng, and N. Riedel, "Discriminant Analysis: A Least Squares Approximation View," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2005.
[22] W. Zheng, L. Zhao, and C. Zou, "An Efficient Algorithm to Solve the Small Sample Size Problem for LDA," *Pattern Recognition,* vol. 37, pp. 1077-1079, 2004.
[23] J. Ye, R. Janardan, C.H. Park, and H. Park, "An Optimization Criterion for Generalized Discriminant Analysis on Undersample Problems," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 26, no. 8, pp. 982-994, Aug. 2004.