Shared Feature Extraction for Nearest Neighbor Face Recognition

David Masip and Jordi Vitrià

***Abstract*—In this paper, we propose a new supervised linear feature extraction technique for multiclass classification problems that is specially suited to the nearest neighbor classifier (NN). The problem of finding the optimal linear projection matrix is defined as a classification problem and the Adaboost algorithm is used to compute it in an iterative way. This strategy allows the introduction of a multitask learning (MTL) criterion in the method and results in a solution that makes no assumptions about the data distribution and that is specially appropriated to solve the small sample size problem. The performance of the method is illustrated by an application to the face recognition problem. The experiments show that the representation obtained following the multitask approach improves the classic feature extraction algorithms when using the NN classifier, especially when we have a few examples from each class.**

***Index Terms*—Face recognition, feature extraction, multitask learning (MTL), nearest neighbor classification (NN), small sample size problem.**

I. Introduction

THE integration of computers in our everyday life is increasing every day as technology evolves, making feasible new applications deal with automatic face classification problems; among them we can find face recognition applied to security, biometrics, and design of more user-friendly interfaces. Face images are captured as high-dimensional feature vectors, being usually a necessary dimensionality reduction process.

According to the literature, the dimensionality reduction techniques can be classified in two categories: feature selection and feature extraction. In the feature selection [1]–[4] approach, only a subset from the original feature vector is preserved. In feature extraction, the original features are combined or transformed into the new extracted features. In this paper, we will deal with linear feature extraction methods, considering the feature selection problem as a special case of feature extraction where the selected features have coefficient 1 in the projection matrix $\mathbf{P}$, and 0 in the other features.

Classification algorithms receive as an input a set of training samples each one represented as a feature vector. In the statistical pattern recognition literature, we can find two important reasons to reduce the number of features [5]: 1) alleviate the curse of dimensionality problem, improving the parameter estimation of the classifier [6] and 2) reduce the storage and computational needs. These advantages are crucial in face recognition applications, where the feature extraction process can mitigate the effect of the noise present in natural images, and also in order to find invariant characteristics of the individual, making the later classification step more robust to changes in illumination or partial occlusions.

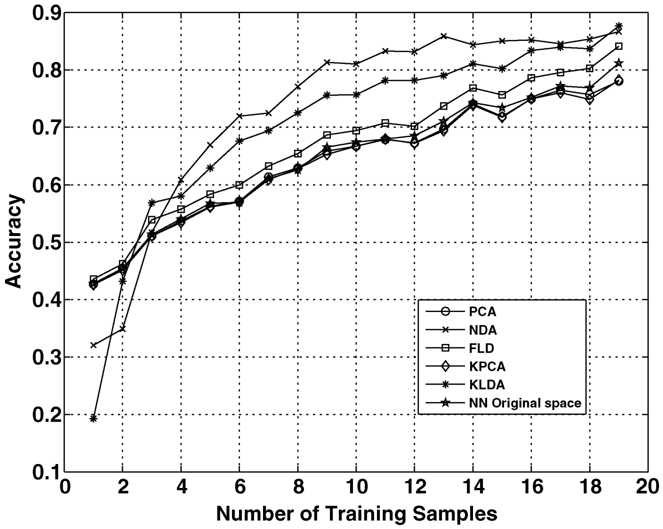


Fig. 1. Example of classification accuracy as a function of the number of training samples using the NN classifier and some of the state-of-the-art feature extraction algorithms. The ARFace database has been used for this experiment.

Typically, in face recognition problems, the number of images from each class is considerably limited: only one or two faces can be acquired from each person. Under this assumption, most of the traditional feature extraction methods suffer a significant performance drop. Fig. 1 shows an example of the evolution of the classification accuracy when applying different feature extraction algorithms with different number of training samples. This phenomenon is known as the *small sample size problem*.

In addition, in face recognition problems, the dimensionality of the input space is rather large. In this context, many statistical classifiers fail in the density estimation task, making the nonparametric classifiers a good alternative. In this paper, we focus on the nearest neighbor classification rule (NN), which is one of the most used in spite of its simplicity. Assuming the NN classifier, the main motivation of this work is to provide a multiclass feature extraction method that performs no specific distribution assumptions on the data, and that can improve the results of the NN classification even when only a few samples per class are available.

The first contribution of this paper is to propose a new iterative feature extraction method for multiclass classification under

Manuscript received July 25, 2006; revised May 11, 2007; accepted September 4, 2007. This work was supported by the Ministerio de Educación y Ciencia (MEC) under Grant TIN2006-15308-C02-01, Spain.

D. Masip is with the Universitat Oberta de Catalunya, Barcelona 08018, Spain (e-mail: dmasipr@uoc.edu; davidm@maia.ub.es).

J. Vitrià is with the Computer Vision Center, Computer Science Department, Universitat Autònoma de Barcelona, Barcelona 08193, Spain (e-mail: jordi@cvc.uab.es).

Digital Object Identifier 10.1109/TNN.2007.911742

1) Calculate the criterion matrix $\mathcal{M}$.
2) Compute the eigenvalues and the corresponding eigenvectors of $\mathcal{M}$
3) Sort the eigenvalues of $\mathcal{M}$ and take the eigenvectors corresponding to the M largest eigenvalues to build the projection matrix.

Fig. 2. General procedure for linear feature extraction.

the NN classification rule. Usually, feature extraction algorithms have been used as a previous and independent step when training a classifier, but recent studies show that this process can be considered during classifier construction, resulting in very strong classifiers [7].

In this paper, we propose to reverse the problem and not to use feature extraction for classification but classification for extracting features, that is, to define the feature extraction as a classification problem. This approach is based on a previous work [8], which was restricted to two-class problems. In [9], it was experimentally shown that Adaboost-based feature extraction is specially suited to high-dimensional data, which is the case of face recognition problems. Our proposal is to extend this method to the multiclass case to allow full face recognition using feature extraction and the NN classifier. We make no specific assumptions on the class distribution of the data, and incrementally build the supervised projection matrix that enhances the most the extra-class separability at each step. In order to extend the original algorithm to the multiclass case, we use the gentleboost algorithm [10] with the Adaboost.MH cost function [11].

The second contribution of this paper is the use of a multitask learning (MTL) approach in the feature extraction task [12]. The MTL approach learns from different tasks in parallel, obtaining improved classification rates given that the "knowledge" learned for one task can help in learning about the other tasks. Inductive transfer [13] and MTL disciplines have been subject of research during the recent years, and they have been focused on classification tasks.

The term MTL was first introduced by Caruana in [12]. His work on predicting the best possible steering direction on road images is one of the first applications of MTL on the computer vision field. Intrator and Edelman [14] also applied the MTL approach to image classification tasks. Torralba *et al.* [15] adapted the multitask approach to the Adaboost classifier. They proposed a joint boosting algorithm for an object detection problem with background and multiple objects involved. Their results show an improvement on the accuracies, obtaining faster classifiers, and also requiring less training data.

More concretely, in this paper, we propose to introduce the MTL approach into the feature extraction task. A linear projection matrix is obtained selecting 1-D feature extractors shared among the different classification problems, using the multiclass extension of the Adaboost-based feature extraction algorithm [9]. This approach is applied to a face recognition problem where we use few training images from each class. Under these assumptions, the performed experiments show improved accuracies in comparison to the classic feature extraction techniques using the NN classifier. Moreover, given that our approach makes no specific density assumptions relatively to the data distribution and supposing that the initial data set is diverse enough, we have the following two hypothesis: 1) the learned features that are shared among existing classes will be also useful when new classes are added to the system and 2) classification performance should not degrade after the addition of a new person in the biometric system.

In Section II, we review some of the state-of-the-art linear feature extraction algorithms applied to the NN face recognition, and the importance of the small sample size problem. Section III introduces the MTL paradigm, and the proposed shared boosting feature extraction algorithm. Then, some experimental results on three face databases are shown, and finally, we end with the conclusions and future work.

II. Machine Learning and Face Recognition

In the last decades, a plethora of face recognition methods have been developed. A first approach to a high-level categorization [16] divides the algorithms into the following: 1) *holistic methods* where the whole face region is used as a raw input to the recognition system and 2) *structural matching methods*, where local features location or statistics are computed in certain parts of the face, and their geometry is used in a structural classifier. Also, hybrid methods have been developed. In contrast to the case of object detection and general object recognition, most of the actual state-of-the-art face recognition algorithms use the appearance-based holistic approach [17], taking a previous feature extraction step to reduce the data dimensionality and preserve the most discriminant information.

These strategies have been shown to be successful in the literature, assuming that enough training samples for each face are available. Nevertheless, in many of the real-world applications such as law enforcement or document identification, the number of training samples is reduced, being that the global performance is considerably diminished [18]. In this section, we review some of the most often used feature extraction techniques in the face recognition literature, and the last contributions made to the small sample size problem.

A. Feature Extraction

Linear feature extraction methods can be expressed as a matrix product that performs the mixture and selection of the original features. Given a data set $\mathbf{X} = [\mathbf{x}_1^T, \mathbf{x}_2^T, \ldots \mathbf{x}_N^T]$, the goal of the linear feature extraction is to find a transformation from the original to a reduced space $\mathcal{F} : \Re^D \rightarrow \Re^M$, usually $(M \ll D)$, that can be expressed as $\mathbf{y_i} = \mathbf{P}\mathbf{x}_i$. There are plenty of linear feature extraction methods, which can be classified into unsupervised (when there is no prior information about the data class membership) and supervised.

The best known linear unsupervised feature extraction technique is the principal component analysis (PCA) [19] that optimizes the linear projection minimizing the mean square reconstruction error criterion. Gaussian distribution is assumed when finding the axis that retains the maximum variance. Fig. 2 shows the general procedure to find the linear projection $\mathbf{P}$. In the case of PCA, the covariance matrix is used as the criterion matrix $\mathcal{M}$.

On the other hand, independent component analysis (ICA) [20]–[23] performs no second-order assumptions on the data, obtaining a linear combination of independent sources. Other feature extraction techniques have been developed for special cases such as nonnegative matrix factorization (NMF) [24]–[27], which adds the positivity constraint to the projection matrix and extracted features, obtaining sparse feature sets.

Supervised feature extraction [28] methods use the labels of the data $(C_1, \ldots, C_K)$ to find the most discriminatory features. The best known supervised method is Fisher linear discriminant analysis (FLD), which maximizes the separability measure

$$\mathcal{J} = \operatorname{tr}\left(\left(\mathbf{P}\mathbf{S}_W^{-1}\mathbf{P}^T\right)\left(\mathbf{P}\mathbf{S}_B\mathbf{P}^T\right)\right). \tag{1}$$

The between-class scatter $(\mathbf{S}_B)$ and the within-class $(\mathbf{S}_W)$ matrices are defined as

$$\mathbf{S}_W = \frac{1}{K}\sum_{k=1}^{K}\mathbf{S}_k \tag{2}$$

$$\mathbf{S}_B = \frac{1}{K}\sum_{k=1}^{K}(\mathbf{m}_k - \mathbf{m}_0)(\mathbf{m}_k - \mathbf{m}_0)^T \tag{3}$$

where $\mathbf{S}_k$ is the class-conditional covariance matrix for the class C_k, $\mathbf{m}_k$ is the class-conditional sample mean, and $\mathbf{m}_0$ is the global sample mean. The linear projection is found using $\mathbf{S}_W^{-1}\mathbf{S}_B$ as $\mathcal{M}$ in Fig. 2. There is a Gaussian assumption on the class distribution, being that the algorithm is blind beyond the second-order statistics.

Fukunaga and Mantock [29], [30] propose a method to overcome the main drawbacks of an FLD, the nonparametric discriminant analysis (NDA). The between-class scatter matrix is computed as an average of local covariance matrices. Similarly, Bressan and Vitrià [31] introduced a nonparametric approach for the within-class scatter matrix. More recently, Loog and Duin [32] have introduced an extension for the FLD formulation for the case of heteroscedastic data; they use the Chernoff criterion achieving improved results.

Each of these techniques has its own limitations, depending on the assumption it makes. For example, FLD projection is $\mathcal{F} : \Re^D \rightarrow \Re^M$, $(M = K - 1)$, where K is the number of different classes. This is a very restrictive condition when dealing with high-dimensional data. Moreover, its assumption about Gaussian distribution of data makes this technique not suitable for multimodal class distributions.

Recently, a new linear feature extraction methodology to extract discriminant features has been proposed [9]. The method is based on the iterative computation of the projection matrix by following a boosting strategy. The proposal tries to overcome the two main drawbacks of the classic discriminant analysis: the limitation about feature dimensionality and the specific density assumptions on the data. The algorithm incrementally finds the projection matrix using a modified Adaboost algorithm [33], [34]. At each boosting step, a pool of linear projections from the original to the 1-D space is generated. The best projection is selected according to the classification error using the Adaboost weights at the current step. The evaluation of the best projection is performed by constructing a decision stumps classifier on the projected training data [10]. The Adaboost weights are adjusted according to the classification error at each step. Each single 1-D projection is selected to be a column vector of the final projection matrix that performs the linear feature extraction. The Adaboost weight adjustment ensures diverse enough projections, obtaining a projection matrix that does not make specific density assumptions on the training data distributions. The experiments performed in previous works [9] show improved accuracies in high-dimensional data sets, where usually the Gaussian assumptions are not met. Nevertheless, the method is limited to the two-class problem case, not being suitable for general face recognition.

B. *Face Recognition and Small Sample Size Problem*

Appearance-based methods have been the dominant techniques in the face recognition field since the 1990s. Nevertheless, the performance of this methods is heavily affected by the number of training samples for each face [35]. The proposed solutions to this problem can be grouped into the following three categories.

- Methods based on regularizing the problem using an algebraic approach or an unsupervised preprocess of the data [36], [37]. In addition, kernel methods have been applied to the problem [38].
- Methods that use *a priori* information of the problem to synthetically increase the number of samples [39]–[41].
- Methods that propose to share the knowledge among similar problems to obtain more robust decision rules. This strategy has been successfully applied in the generic object recognition field [15], [42]–[45] but, to our knowledge, it has been only used in the face classification field in [46]. This method includes *a priori* information on the intraclass scatter matrix by adding faces from other subjects, assuming that human faces share a common structure. In the second step, the method applies a Fisher criterion to find the projection matrix.

In Section III, we propose a new method that follows the sharing knowledge strategy, the multitask boosted discriminant projections (MBDP), preserving the advantages of the linear discriminant feature extraction methods successfully used in face recognition.

III. Boosted Shared Linear Discriminant Features

As explained in Section II, linear discriminant analysis techniques usually maximize a specific criterion based on some assumptions made on the training data. Nevertheless, the boosted feature extraction method proposed in [9] makes no statistical assumptions on the data, being exclusively a *classifier-driven* feature extraction method. We propose to extend this method in two directions: define a formulation that can afford the multiclass case and allow the use of shared features to increase the reliability in presence of the small sample size problem.

To introduce the sharing information procedure in the feature extraction framework, we follow the MTL paradigm. The MTL approach is a new classifying methodology based on jointly training multiple related problems and named tasks, and it takes advantage of its relationship to improve the learning process of

```
1) Given the matrix $\mathbf{X}$ containing data samples $\mathbf{x}_i$, and the vector $\mathbf{c}$ with the corresponding labels $c_i \in \{C_1, C_2, \dots, C_K\}$ $(i = 1 \dots N)$
2) Initialize a set of weights: $\mathbf{W}_i^c(1) = 1$.
3) For $t = 1 \dots M$:
   a) For all the possible ways of grouping the classes in binary problems $n = 1, \dots, 2^K$:
      i) Generate $P$ projections from the original space to an $1-$dimensional subspace.
      ii) Project the training data into the space defined by the projections $\mathbf{p}_1, \dots, \mathbf{p}_P$.
      iii) Learn the shared regression stumps classifier on the projected data, obtaining the hypothesis:
           $h_t^n(\mathbf{x}_i, c) = \begin{cases} \alpha\delta(\mathbf{x}_i^j > \theta) + \rho, & \text{when } C_i \in \text{Positive(n)} \\ k^c, & \text{when } C_i \notin \text{Positive(n)} \end{cases}$   (4)
           Denoting by $\mathbf{x}_i^j$ the j-th feature of the projected sample $\mathbf{x}_i$, where the decision stumps classifier obtains the maximum accuracy on the training data.
      iv) Compute the weighted error for the class grouping as:
           $Err_p(n) = \sum_{c=1}^{K} \sum_{i=1}^{N} \mathbf{W}_i^c (b_i^c - h_t^n(\mathbf{x}_i, c))^2.$   (5)
      where $b_i^c \in \{-1, +1\}$ is the binary class label assigned to the class $C_i$ in the n-th binary grouping.
   b) Find the binary grouping of classes m with minimum $Err_p$:
           $m = argmin_n Err_p(n)$   (6)
   c) Update the data weights:
           $\mathbf{W}_i^c(t+1) = \mathbf{W}_i^c(t) exp^{-b_i^c h_t^m(\mathbf{x}_i, c)}, \quad i = 1, \dots, N.$   (7)
   d) Find the projection $\mathbf{m}$ associated to the feature where the decision stumps obtained the minimum training error.
   e) Store $\mathbf{m}$ as the $t$-th projection in the projection matrix $\mathbf{P}$.
4) Output the projection matrix $\mathbf{P}$, built using the vectors selected at each boosting step as columns.
```

Fig. 3. Scheme of the shared boosted feature extraction algorithm (MBDP).

the individual classification tasks. The previous works on MTL show interesting improvements at two different levels: the accuracies of the methods increase (parallel transfer) when problems are jointly trained, and the number of samples needed to train reliable classifiers decreases (sequential transfer). The advantages of MTL have been experimentally validated in the first works of Thrun [47] and Baxter [48]. The following two different approaches to MTL can be identified in the recent literature:

- A functional approach, where the tasks share the hypothesis space [13], [49]. A typical example are the multitask support vector machines (M-SVM) [50], [51].
- A structural approach, where the representation of the data is supposed to share a common generative process that can be used in the hypothesis selection [12], [14], [52].

The applications of MTL on computer vision are still not deeply explored, but it seems *a priori* that the second approach can be the most suitable for the face recognition tasks. We consider the case where only a few examples from each class are available, which usually is the case of face recognition problems, and most of them will suffer from illumination changes. The MTL applied to feature extraction can be a useful tool to focus the projection vectors on the general recognition task, discarding the intravariations due to illumination. Torralba *et al.* [15] followed this approach but applied it to a direct classification scheme.

A. MBDP Algorithm

The general algorithm of the proposed method is shown in Fig. 3. In this section, we explain the details of the final implementation.

As in the two-class version, the algorithm takes as input the N training samples $\mathbf{X}$ $(D \times N)$ and the corresponding labels $C_i \in \{C_1, \dots, C_K\}$. We perform M boosting rounds in order to obtain one 1-D projection at each step, being that the output of the algorithm is the projection matrix $\mathbf{P}$ $(D \times M)$ made by placing the local projections at each step as columns. At each boosting step, the multiclass problem is transformed into multiple binary problems by grouping the classes in a positive and a negative cluster. The algorithm should find the best grouping according to the training error. Nevertheless, the number of possible groupings grows exponentially with the number of classes $O(2^K)$. This fact makes the problem computationally unfeasible in face recognition tasks where usually the number of classes is large enough. In this paper, we have followed the best first search $O(K^2)$ as in [15], grouping the classes as follows.

1) Train a degenerated decision tree classifier [15] using a single class as positive.
2) Select the class with minimum classification error as the initial positive cluster.
3) For the remaining $K-1$ classes
   - train a classifier using the previous positive cluster but adding another class from the negative cluster;
   - add to the previous positive cluster the class from the negative cluster only if the joint selection improves the previous classification error.

At each boosting step, a new grouping may be selected in order to reduce the weighted general error. The class grouping allows to share knowledge among classes, being that the extracted features are less specific. For each possible grouping, a set of candidate $\mathbf{p}_1, \ldots, \mathbf{p}_P$ projections is trained in the general case. The training data is projected according to $\mathbf{p_i}$, and a decision stumps classifier is trained in order to construct the classification hypothesis $h_t^n(\mathbf{x}_i, c)$ for the grouping n. The regression stumps parameters are found as in [15]

$$\rho = \frac{\sum_{c \in \text{Positive}(n)} \sum_i \mathbf{W}_i^c b_i^c \delta\left(\mathbf{x}_i^j \leq \theta\right)}{\sum_{c \in \text{Positive}(n)} \sum_i \mathbf{W}_i^c \delta\left(\mathbf{x}_i^j \leq \theta\right)} \tag{8}$$

$$\alpha + \rho = \frac{\sum_{c \in \text{Positive}(n)} \sum_i \mathbf{W}_i^c b_i^c \delta\left(\mathbf{x}_i^j > \theta\right)}{\sum_{c \in \text{Positive}(n)} \sum_i \mathbf{W}_i^c \delta\left(\mathbf{x}_i^j > \theta\right)} \tag{9}$$

$$k^c = \frac{\sum_i \mathbf{W}_i^c b_i^c}{\sum_i \mathbf{W}_i^c}, \qquad \text{if } c \notin \text{Positive}(n) \tag{10}$$

where k^c acts as a constant to prevent a class from being more frequently selected due to an imbalanced training set.

The class grouping with lowest training error is selected, and the weights $\mathbf{W}$ are adjusted according to the multiclass gentleboost algorithm. The projection vector that generated the selected minimum error feature in the decisions stumps classification is selected as the next column of the projection matrix $\mathbf{P}$. Notice that the partial classifiers are discarded, and the output of the algorithm is a single projection matrix; the result of the ensemble on the training set is discarded.

Step 3ai) in Fig. 3 defines how the local projections of each boosting step are generated. Depending on the algorithm used to obtain these projections different methods can be derived from the original solution. Previous works suggest the use of pseudorandom or local projections, or simply to sample the training set and use a classic discriminant analysis algorithm on the binary grouping (the reader can find more details on the stability of this approaches in two-class problems in [9]). In this paper, we use a single FLD projection obtained from the binary problem, although many variations of the algorithm could be used given the proper feature extraction algorithm.

IV. Experiments

In this paper, we have used the proposed feature extraction method on a face recognition problem, where only a few training images are available from each person. We have used the NN as a classification rule on the extracted features. To validate our proposal, we have performed a set of experiments using a representative amount of feature extraction techniques: the classic FLD, the NDA, and the unsupervised PCA. In addition, the more recent kernel PCA (KPCA) algorithm was used in the comparison, and the kernel LDA (KLDA) method from Lu *et al.* [38] has been used as the reference technique from the recent literature. The goal of the experimental section is to verify whether our MBDP proposed method outperforms the classic and the state-of-the-art feature extraction techniques. Moreover the classification accuracies using directly the NN rule on the original space have been used for comparison, to validate the usefulness of the feature extraction techniques as a previous step to the NN classification.

TABLE I
Summary of the Main Characteristics of the Testing Data Sets

Database	Dimensionality	Number of samples	Number of classes
FRGC	2272 × 1074	3772	275
ARFace	768 × 576	4000	126
ORL	92 × 112	400	40

The experiments performed use three different databases: the face recognition grand challenge data set (FRGC) [53], the AR face database [54], and the ORL database [55]. The original FRGC 1.0 database has 3772 samples from 275 different subjects, between four and 32 high-quality images (typically 2272 pixels × 1704 pixels) per person can be found depending on the subject. The AR face database contains over 4000 images acquired in two sessions under different conditions. For each person and session, we have one neutral image, one smiling, one anger, one screaming, three images with strong light artifacts, three images with occlusions due to sun glasses and illumination changes, and three images with scarf occlusions and illumination changes. The ORL database contains ten images (92 pixels × 112 pixels) from 40 distinct subjects (400 samples in total) acquired in the AT&T Laboratories, Cambridge, U.K. Table I summarizes the main characteristics of each data set.

In all the cases, a previous preprocessing has been performed, given the illumination variability of most of the samples.

A. Preprocessing

To extract the internal features of the face image, the following normalizing steps have been followed.

- The images were converted from RGB to grayscale.
- A light normalization based on anisotropic filtering on the original image resolution was performed [56].
- Then, we performed a geometric normalization using the coordinates of each eye. Images have been rotated and scaled according to the intereye distance, in such a way that the center pixel of each eye coincided in all the images. The samples were then cropped obtaining a 37 × 33 thumbnail, so only the internal region of the faces was preserved. No hair information was used in the classification task.
- In order to avoid the remaining data in the external face region, the thumbnails were masked using an elliptical mask.
- A histogram equalization in the nonmasked pixels as done in [57] were performed.
- Finally, the pixels have been normalized to have zero mean and standard deviation of one.

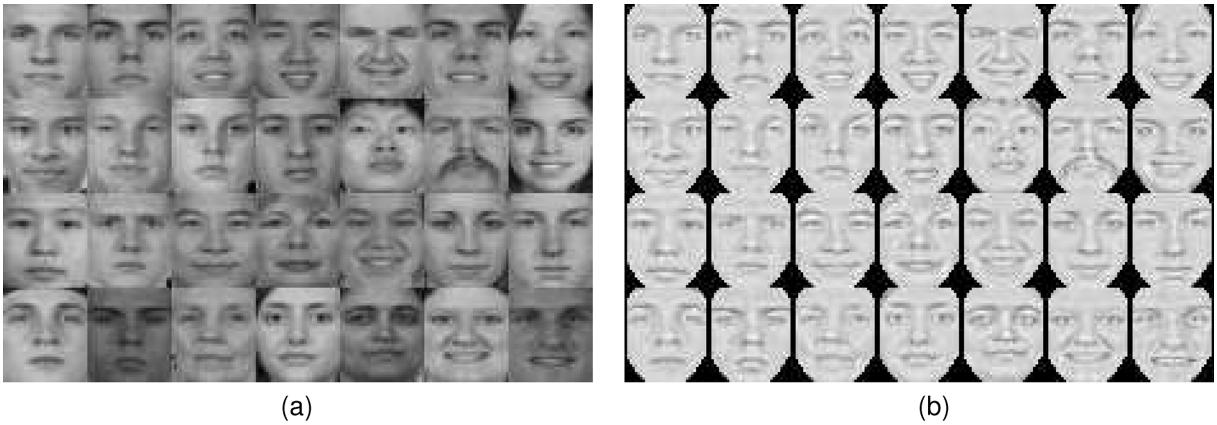


Fig. 4. Example of some face images used in the experiments. (a) Original faces. (b) Normalized version of the faces.

TABLE II
BEST MEAN RECOGNITION RATES WITH THE CONFIDENCE INTERVALS AND THE DIMENSIONALITY WHERE THEY ARE ACHIEVED

Algorithm	FRGC	AR face	ORL	Average Ranks
PCA	$0.8446 \pm 0.014(98)$	$0.7325 \pm 0.023(99)$	$0.8162 \pm 0.011(79)$	4.7
NDA	$0.8132 \pm 0.019(45)$	$0.7100 \pm 0.020(30)$	$0.7948 \pm 0.025(39)$	7.0
FLD	$0.8659 \pm 0.015(49)$	$0.7645 \pm 0.018(49)$	$\mathbf{0.8459}* \pm 0.009(39)$	2.3
KPCA	$0.8495 \pm 0.015(97)$	$0.7325 \pm 0.023(99)$	$0.8181 \pm 0.012(79)$	4.3
KLDA	$\mathbf{0.8797} \pm 0.018(49)$	$0.7850 \pm 0.016(48)$	$0.8284 \pm 0.008(39)$	2.3
MBDP	$\mathbf{0.9109}* \pm 0.018(100)$	$\mathbf{0.8566}* \pm 0.002(93)$	$\mathbf{0.8290} \pm 0.012(93)$	1.3
NN	$0.8446 \pm 0.045(-)$	$0.7325 \pm 0.072(-)$	$\mathbf{0.8081} \pm 0.045(-)$	6.0

Exceptionally, in the case of the ORL face database, the original 92 × 112 samples have been used omitting the scaling step. The reduced nature of the ORL data set computationally allows us to test the different feature extraction methods in the original high-dimensional subspace.

Fig. 4 shows some examples of the face images and the normalized version used to train the feature extraction task.

B. *Classification Results*

The experiments have been repeated 100 times and only the mean of the experiments is shown. At each repetition of the algorithms, 50 out of the total number of classes from each database are randomly taken (40 in the case of the ORL data set). The six feature extraction algorithms are trained using only two samples from each class (the minimum number of samples needed for discriminant analysis techniques as shown in Section II). The rest of the samples were used for testing. The feature extraction algorithms have been trained to extract up to 100 features (although there are no theoretical bounds in the subspace dimensionality obtained using the shared boosted method). Using the classic discriminant analysis techniques, the number of features is limited to 49. In the NDA approach, a previous dimensionality reduction using PCA preserving the 95% of the variance has been performed. It has been shown that this previous step improves significantly the accuracies obtained, as the NDA can be considerably affected by the noise components [31].

In Table II, we show the mean recognition rates using each method and the best dimensionality where they are achieved. Also, the 95% confidence intervals are shown. The best performing algorithm is marked with the "$*$" for each data set, and the algorithms whose confidence intervals overlap with the best performing method are marked in bold face. Ranks have been assigned to each method, where rank 1 is the best performing method, and rank 7 is the worst. The last column of the table shows the average ranks along the three data sets of the different feature extraction algorithms.

The proposed MBDP method obtains the best average rank, being the best method in two out of the three data sets. Moreover, the three supervised dimensionality reduction techniques (MBDP, KLDA, and LDA) seem to outperform the rest of the feature extraction techniques. To statistically test the stability of these results, the nonparametric Friedman test variant of the analysis of variance (ANOVA) has been performed. For more details on the use of appropriate statistical test on classification performances the reader can see [58]. Briefly, the Friedman tests goal is to reject the null hypothesis that advocates that the algorithms are equivalent, and the performance differences observed can be considered merely random. First, the Friedman statistic is computed as

$$\chi_F^2 = \frac{12N}{k(k+1)} \left[\sum_j R_j^2 - \frac{k(k+1)^2}{4} \right] \tag{11}$$

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1) - \chi_F^2} \tag{12}$$

where N is the number of data sets, k is the number of classification methods, and R_j is the average rank of the algorithm j. The F_F is distributed according to an F-distribution with $k-1$ and $(\mathrm{k}-1)(N-1)$ degrees of freedom. According to a prefixed confidence p-value, a critical value of the F_F statistic can be found on the F-distribution table. In our case, the tabulated critical value for $F(k-1,(K-1,N-1)) = F(6,12) = 4.821$

TABLE III
RESULTS OF THE BONFERRONI–DUNN *Post Hoc* STATISTICAL TEST PERFORMED. THE MBDP ALGORITHM IS USED AS A CONTROL CLASSIFIER FOR COMPARISON. THE TERM SE STANDS FOR $\sqrt{(k(k+1)/6N)}$

i	Algorithm	$z = (R_0 - R_i)/SE)$	p-value	$\alpha/(k-i)$
1	NDA	3.2127	0.0013	0.0083
2	NN	2.6458	0.0081	0.0100
3	PCA	1.8898	0.0588	0.0125
4	KPCA	1.7008	0.0890	0.0167
5	FLD	0.5669	0.5708	0.0250
6	KLDA	0.5669	0.5708	0.0500

TABLE IV
RESULTS OF THE BONFERRONI–DUNN *Post Hoc* STATISTICAL TEST PERFORMED USING THE NN CLASSIFICATION ON THE ORIGINAL SPACE AS A CONTROL METHOD FOR COMPARISON

i	Algorithm	$z = (R_0 - R_i)/SE)$	p-value	$\alpha/(k-i)$
1	MBDP	2.6458	0.0081	0.0083
2	FLD	2.0788	0.0376	0.0100
3	KLDA	2.0788	0.0376	0.0125
4	KPCA	0.9449	0.3447	0.0167
5	PCA	0.7559	0.4497	0.0250
6	NDA	0.5669	0.5708	0.0500

is lower than the $F_F = 29.5$ obtained, therefore we reject the null hypothesis, reporting significance on the results.

Moreover, once the statistical significance of the results is assured, we seek a statistical evidence to check whether an algorithm outperforms the rest. For this propose, a *post hoc* significance test has been performed, using the proposed MBDP method as a control classifier. To compare each classifier with the MBDP, the z-statistic has been used

$$z = \frac{(R_j - R_j)}{\sqrt{\left(\frac{k(k+1)}{6N}\right)}} \tag{13}$$

which is normally distributed. The ordered p-values $p_1 \leq p_2 \leq \ldots \leq p_{k-1}$ of each comparison have been computed in Table III, following the Bonferroni–Dunn procedure, and the critical values for the $\alpha = 0.05$ confidence are also shown. Only the p-values under the $\alpha/(k-i)$ critical threshold can be rejected, assuring statistical significance. The reader can find more details on this procedure in [58]. Using this restrictive statistic, we can assure that MBDP improves the performance of the NN on the original space and the subspace defined by NDA. No more inferences can performed, being that the *post hoc* test is not precise enough to detect other nonoverlapping significant differences among other methods.

Nevertheless, additional information can be extracted from these results. The main goal of any feature extraction algorithm is to improve the classification results of the posterior classifier by reducing the complexity of the input data. We propose the hypothesis that the six methods used in the test improve the NN rule directly on the original space, and test its significance by performing a second *post hoc* test following the same Bonferroni–Dunn procedure, using the NN on the original space as a control method. Table IV summarizes the p-values obtained. The hypothesis can be assured only in the first case (using the MBDP proposed).

Fig. 5 shows the mean accuracies as a function of the number of extracted features. Notice that different subspace dimensionality is plotted depending on the method, and the classic discriminant approaches only obtain dimensionality subspaces of up to 49 (39 in the case of the ORL data set). Nevertheless, our boosted feature extraction approach is not upper bounded. Moreover, to facilitate a better interpretation of the results, as a horizontal line, we plot the accuracy of the NN classifier applied directly in the original space.

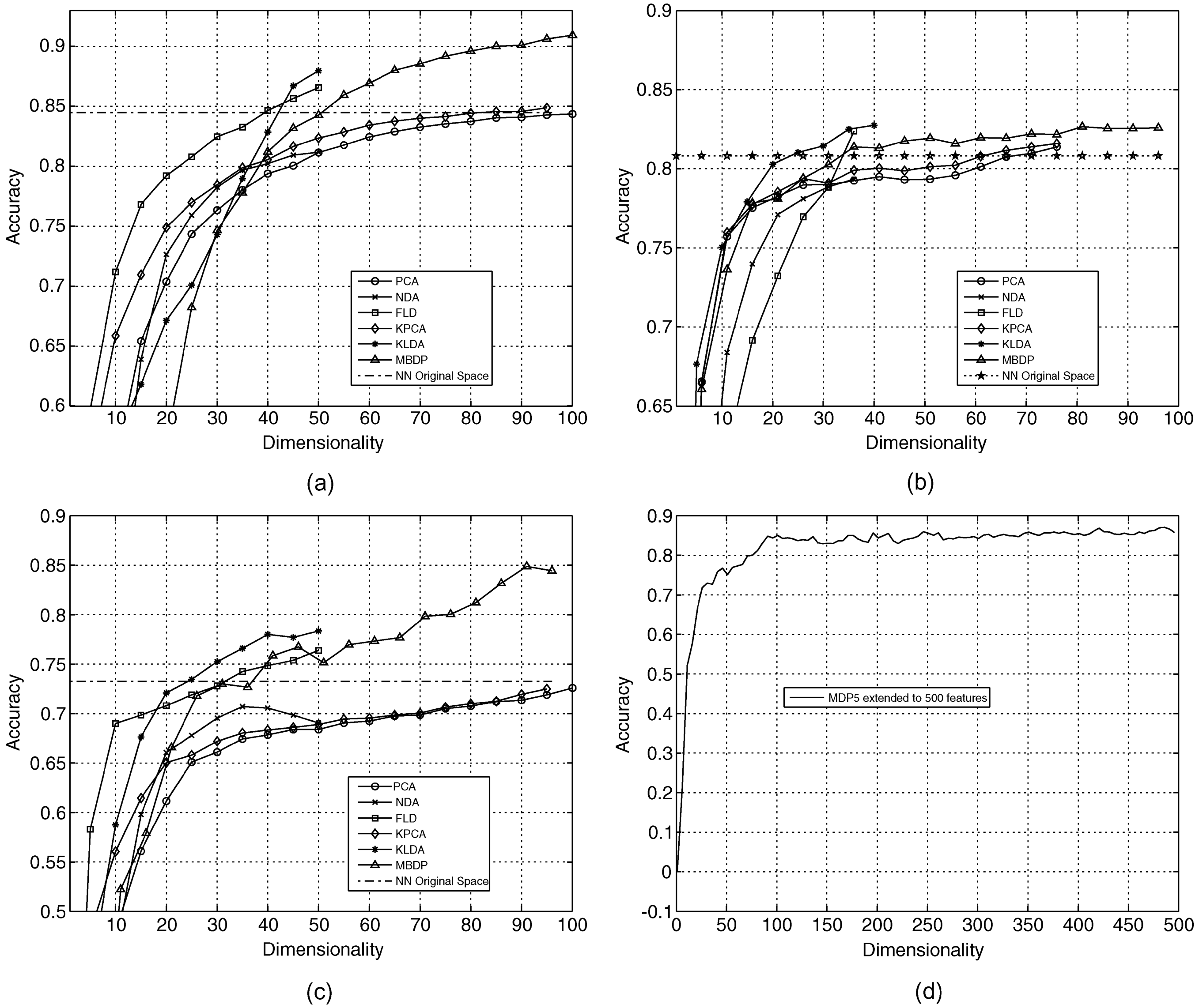
In the case of the FRGC data set, the proposed MBDP feature extraction method outperforms the other techniques from the 65th feature extracted. As it was expected, our method fares worse than FLD in the lower dimensional subspaces, given that the first features extracted are a direct application of FLD on a binary relabeled data set. Notice also that the performance of the nonsupervised PCA technique is asymptotically close to the one obtained directly using the NN classifier on the original space as the number of extracted features increases. Nevertheless, our approach significantly increases the performance as the dimensionality increases, considerably improving the NN on the original space. This suggests the possibility of the usefulness of our approach in dimensionality augmentation problems.

Similarly, in the AR face case, the proposed method gradually outperforms the other feature extraction techniques, being the best choice as the number of extracted features increases. Notice that accuracy of the shared method using 100 features is 10% higher than directly applying the NN to the original space (using 1221 features). Nevertheless, the accuracy of the shared approach slightly increases as more features are extracted. A second experiment has been performed extracting up to 500 features [Fig. 5(d)], obtaining an accuracy of 0.87 (2% higher).

C. Shared Projections and Class Sampling

The main algorithm proposed in this paper can be computationally demanding in the training phase, while the testing phase has exactly the same computational cost as any other linear feature extraction method. The most costly step is the computation of the best class grouping, given that an $O(K^2)$ combinations of groupings must be analyzed to find the best projection. In problems where the number of classes is large $(K > 1000)$, it is computationally unfeasible to obtain the linear projection in a reasonably bounded amount of time. To solve this problem we propose a modification on the algorithm. The pool of classes available in step 3a) in the general algorithm from Fig. 3 has been limited to a fixed number. At each iteration, we sample K' classes from the original K, and the possible groupings are selected according to these K' classes. At each boosting step, the sampled class pool changes, given that the sampling is performed according to the distribution of the normalized weight of each class on the boosting process. The experiments shown in this paper have been performed with class sampling using $K' = 10$. The same experiment was repeated using the original 50 classes, and the mean accuracies obtained were only 1.7% better.

This modification allows to compute the projection matrix in an approximately constant computational time, independently of the number of classes.



(a)

(b)

(c)

(d)

Fig. 5. Accuracy in the FRGC, AR face, and ORL data sets as a function of the dimensionality reduction of the PCA, NDA, FLD, KPCA, KLDA, and MBDP and the horizontal line that shows the performance of the NN using the original space. In (d) the MBDP has been applied to the AR Fface data base extracting up to 500 features. (a) FRGC data set. (b) ORL data set. (c) AR face data set. (d) MBDP on the AR face data set using up to 500 features.

V. Conclusion

In this paper, we have introduced a new feature extraction algorithm based on applying a modified Adaboost algorithm. The method performs no specific assumptions on the data to classify, and it incorporates the MTL theory extending the previous work to deal with multiclass face recognition problems. The algorithm has been applied to a problem of face recognition from small sample size sets outperforming the classic feature extraction methods based on an analytic discriminant criterion.

The method proposed in this paper obtains an interesting performance in high-dimensional databases, validated using nonparametric statistical tests. This tests show significance in applying the feature extraction method to reduce the complexity of the original space using the NN classifier. Nevertheless, some future works can be done. The diversity on the projection selection in the main step of the algorithm depends on the grouping performed, and it seems the key part of the algorithm. The results as a function of the extracted features indicate that the accuracy of the method increases slower than classic discriminant analysis techniques in the first steps of the algorithm, obtaining a constant grow in the later steps. This fact is related to the diversity of the projections obtained on the first iterations that are only guided for the Adaboost weights. Some explicit diversity measure could be applied in order to speed up the increase on the accuracy and to avoid possible redundant projections.

We also have seen that most of the feature extraction techniques based on eigendecomposition obtain a performance that is asymptotically close to the NN on the original space. However, the accuracy of the proposed method increases as more features are added. This facts suggests the study of dimensionality augmentation instead of the reduction using the boosted approach.

The main drawback of the algorithm is the computational cost in the training phase. We have suggested a modification to reduce this cost by performing a sampling on the class pool, being that the final accuracies are only slightly reduced. Nevertheless, its use in testing time is exactly the same as any other linear feature extraction method.

Also, the MTL model is not completely developed for the face classification task. Other independent tasks could be added to the learning process in order to share more information between the different "subproblems." For example, a task dealing with gender or race recognition would help the recognition problem. The addition of face classification subtasks different from direct person recognition is the most immediate future research line.

References

[1] A. Jain and D. Zongker, "Feature selection: Evaluation, application, and small sample performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 2, pp. 153–158, Feb. 1997.

[2] N. Pal and V. Eluri, "Two efficient connectionist schemes for structure preserving dimensionality reduction," *IEEE Trans. Neural Netw.*, vol. 9, no. 6, pp. 1142–1154, Nov. 1998.

[3] S. Lespinats, M. Verleysen, A. Giron, and G. Fertil, "DD-HDS: A method for visualization and exploration of high-dimensional data," *IEEE Trans. Neural Netw.*, vol. 18, no. 5, pp. 1265–1279, Sep. 2007.

[4] J. M. Leiva-Murillo and A. Artes-Rodriguez, "Maximization of mutual information for supervised linear feature extraction," *IEEE Trans. Neural Netw.*, vol. 18, no. 5, pp. 1433–1441, Sep. 2007.

[5] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4–37, Jan. 2000.

[6] C. M. Bishop, *Neural Networks for Pattern Recognition*. London, U.K.: Oxford Univ. Press, 1995.

[7] J. Lu, K. Plataniotis, A. Venetsanopoulos, and S. Li, "Ensemble-based discriminant learning with boosting for face recognition," *IEEE Trans. Neural Netw.*, vol. 17, no. 1, pp. 166–178, Jan. 2006.

[8] D. Masip and J. Vitria, "Boosted discriminant projections for nearest neighbor classification," *Pattern Recognit.*, vol. 39, no. 2, pp. 164–170, 2006.

[9] D. Masip, L. I. Kuncheva, and J. Vitria, "An ensemble-based method for linear feature extraction for two-class problems," *Pattern Anal. Appl.*, vol. 8, pp. 227–237, 2005.

[10] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: A statistical view of boosting," *Ann. Statist.*, vol. 28, pp. 337–374, 2000.

[11] R. E. Schapire and Y. Singer, "BoosTexter: A boosting-based system for text categorization," *Mach. Learn.*, vol. 39, no. 2/3, pp. 135–168, 2000.

[12] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.

[13] J. Baxter, "A model of inductive bias learning," *J. Mach. Learn. Res.*, vol. 12, pp. 149–198, 2000.

[14] N. Intrator and S. Edelman, "Making a low-dimensional representation suitable for diverse tasks," *Connection Sci.*, vol. 8, pp. 205–224, 1997.

[15] A. Torralba, K. Murphy, and W. Freeman, "Sharing features: Efficient boosting procedures for multiclass object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2004, vol. 2, pp. 762–769.

[16] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surv.*, vol. 35, no. 4, pp. 399–458, 2003.

[17] R. Brunelli and T. Poggio, "Face recognition: Features versus templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 10, pp. 1042–1052, Oct. 1993.

[18] T. Poggio and S. Smale, "The mathematics of learning: Dealing with data," *Amer. Math. Soc. Notice*, vol. 50, no. 5, pp. 537–544, 2003.

[19] J. C. Lv, Z. Yi, and K. K. Tan, "Determination of the number of principal directions in a biologically plausible PCA model," *IEEE Trans. Neural Netw.*, vol. 18, no. 3, pp. 910–916, May 2007.

[20] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001.

[21] P. Comon, "Independent component analysis, a new concept?," *Signal Process.*, vol. 36, no. 3, pp. 287–314, 1994.

[22] J. H. Friedman, "Explanatory projection pursuit," *J. Amer. Statist. Assoc.*, vol. 82, pp. 249–266, 1987.

[23] K. Keun-Chang and W. Pedrycz, "Face recognition using an enhanced independent component analysis approach," *IEEE Trans. Neural Netw.*, vol. 18, no. 2, pp. 530–541, Mar. 2007.

[24] D. Lee and H. S. Seung, "Learning the parts of objects with nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, Jul. 1999.

[25] D. Lee and S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Neural Inf. Process. Syst. Conf.*, 2000, pp. 556–562.

[26] S. Li, X. Hou, and H. Zhang, "Learning spatially localized, parts-based representation," in *Proc. Conf. Comp. Vis. Pattern Recognit.*, Kauai, HI, Dec. 2001, vol. 1, pp. 207–212.

[27] S. Zafeiriou, A. Tefas, I. Buciu, and I. Pitas, "Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification," *IEEE Trans. Neural Netw.*, vol. 17, no. 3, pp. 683–695, May 2007.

[28] M. Kyperountas, A. Tefas, and I. Pitas, "Weighted piecewise LDA for solving the small sample size problem in face verification," *IEEE Trans. Neural Netw.*, vol. 18, no. 2, pp. 506–519, Mar. 2007.

[29] K. Kukunaga and J. Mantock, "Nonparametric discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 5, no. 6, pp. 671–678, Nov. 1983.

[30] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. Boston, MA: Academic, 1990.

[31] M. Bressan and J. Vitria, "Nonparametric discriminant analysis and nearest neighbor classification," *Pattern Recognit. Lett.*, vol. 24, no. 15, pp. 2743–2749, Nov. 2003.

[32] M. Loog and R. P. W. Duin, "Linear dimensionality reduction via a heteroscedastic extension of LDA: The Chernoff criterion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 6, pp. 732–739, Jun. 2004.

[33] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Proc. Int. Conf. Mach. Learn.*, 1996, pp. 148–156.

[34] R. E. Schapire, "A brief introduction to boosting," in *Proc. Int. Joint Conf. Artif. Intell.*, 1999, pp. 1401–1406.

[35] X. Tan, S. Chen, Z.-H. Zhou, and F. Zhang, "Face recognition from a single image per person: A survey," *Pattern Recognit.*, vol. 39, no. 9, pp. 1725–1745, 2006.

[36] R. Huang, Q. Liu, H. Lu, and S. Ma, "Solving the small sample size problem of LDA," in *Proc. 16th Int. Conf. Pattern Recognit.*, Washington, DC, 2002, vol. 3, p. 30029.

[37] P. Howland, J. Wang, and H. Park, "Solving the small sample size problem in face recognition using generalized discriminant analysis," *Pattern Recognit.*, vol. 39, no. 2, pp. 277–287, 2006.

[38] J. Lu, K. Plataniotis, and A. Venetsanopoulos, "Face recognition using kernel direct discriminant analysis algorithms," *IEEE Trans. Neural Netw.*, vol. 14, no. 1, pp. 117–126, Jan. 2003.

[39] D. Beymer and T. Poggio, "Face recognition from one example view," in *Proc. Int. Conf. Comput. Vis.*, 1995, pp. 500–507.

[40] F. D. la Torre, R. Gross, S. Baker, and B. V. K. V. Kumar, "Representational oriented component analysis (ROCA) for face recognition with one sample image per training class," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 266–273.

[41] T. Poggio and T. Vetter, "Recognition and structure from one 2D model view: Observations on prototypes, object classes and symmetries," Tech. Rep. AIM-1347, 1992.

[42] E. Bart and S. Ullman, "Cross-generalization: Learning novel classes from a single example by feature replacement," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Washington, DC, 2005, vol. 1, pp. 672–679.

[43] S. U. Evgeniy Bart, "Single-example learning of novel classes using representation by similarity," in *British Mach. Vis. Conf.*, 2005, pp. 672–679.

[44] F. Li, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," in *Proc. IEEE CVPR Workshop Generative Model Based Vis.*, 2004, pp. 59–70.

[45] K. Levi, M. Fink, and Y. Weiss, "Learning from a small number of training examples by exploiting object categories," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Washington, DC, 2004, vol. 6, p. 96.

[46] J. Wang, K. N. Plataniotis, J. Lu, and A. N. Venetsanopoulos, "On solving the face recognition problem with one training sample per subject," *Pattern Recognit.*, vol. 39, no. 9, pp. 1746–1762, 2006.

[47] S. Thrun and L. Pratt, *Learning to Learn*. Norwell, MA: Kluwer, 1997.

[48] J. Baxter, "Learning internal representations," in *Proc. Workshop Comput. Learn. Theory*, 1995, pp. 311–320.

[49] R. K. Ando and T. Zhang, "A framework for learning predictive structures from multiple tasks and unlabeled data," *J. Mach. Learn. Res.*, vol. 6, pp. 1817–1853, 2005.

[50] T. Jebara, "Multi-task feature and kernel selection for SVMs," in *Proc. 21st Int. Conf. Mach. Learn.*, New York, 2004, pp. 55–62.

[51] T. Evgeniou, C. Micchelli, and M. Pontil, "Learning multiple tasks with kernel methods," *J. Mach. Learn. Res.*, vol. 6, pp. 615–637, 2005.

[52] J. Zhang, Z. Ghahramani, and Y. Yang, "Learning multiple related tasks using latent independent component analysis," in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. Platt, Eds. Cambridge, MA: MIT Press, 2006.

[53] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "The 2005 IEEE workshop on face recognition grand challenge experiments," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Washington, DC, 2005, pp. 947–954.

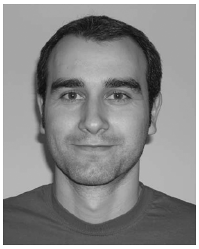
[54] A. Martinez and R. Benavente, "The AR face database," Computer Vision Center, Universitat Autonoma de Barcelona, Barcelona, Spain, Tech. Rep. 24, Jun. 1998.

[55] The ORL Face Database at the AT&T (Olivetti) Research Laboratory [Online]. Available: http://www.uk.research.att.com/facedatabase.html 1992

[56] A. Pujol, "Contributions to shape and texture face similarity measurement," Ph.D. dissertation, Comp. Sci. Dept., Universitat Autonoma de Barcelona, Barcelona, Spain, 2001.

[57] S. A. Rizvi, P. J. Phillips, and H. Moon, "The FERET verification testing protocol for face recognition algorithms," in *Proc. 3rd. Int. Conf. Face Gesture Recognit.*, Washington, DC, 1998, pp. 48–54.

[58] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006.



David Masip received the Ph.D. degree in computer science from the Computer Vision Center (CVC), Universitat Autònoma de Barcelona (UAB), Barcelona, Spain, in 2005.

He has authored more than 20 scientific papers. His research interests concern the development of algorithms of feature extraction, pattern recognition, face classification, and data mining.

Jordi Vitrià received the Ph.D. degree from the Universitat Autònoma de Barcelona (UAB), Barcelona, Spain, for his work on mathematical morphology in 1990.

He joined the Computer Science Department, UAB, where he became an Associate Professor in 1991. He is the author of more than 40 scientific publications and one book. His research interests include machine learning, pattern recognition, and visual object recognition.