ELSEVIER

Signal Processing: *Image Communication* 17 (2002) 657–673

SIGNAL PROCESSING: IMAGE COMMUNICATION

www.elsevier.com/locate/image

A solution for facial expression representation and recognition

S. Dubuisson, F. Davoine, M. Masson*

Laboratory Heudiasyc, U.T.C., B.P. 20529. F-60205 Compiègne, France

Abstract

The design of a recognition system requires careful attention to pattern representation and classifier design. Some statistical approaches choose those features, in a d-dimensional initial space, which allow sample vectors belonging to different categories to occupy compact and disjoint regions in a low-dimensional subspace. The effectiveness of the representation subspace is then determined by how well samples from different classes can be separated. In this paper, we propose a feature selection process that sorts the principal components, generated by principal component analysis, in the order of their importance to solve a specific recognition task. This method provides a low-dimensional representation subspace which has been optimized to improve the classification accuracy. We focus on the problem of facial expression recognition to demonstrate this technique. We also propose a decision tree-based classifier that provides a ''coarse-to-fine'' classification of new samples by successive projections onto more and more precise representation subspaces. Results confirm, first, that the choice of the representation strongly influences the classification results, second that a classifier has to be designed for a specific representation.

Keywords: Facial expression recognition; Dimensionality reduction; Feature selection and extraction; Classifier design

1. Introduction

During the last decade, computers have been equipped with new functions and input/output devices: future human–machine (HM) environments will integrate multimodal information issuing from signs revealing the emotional states of their users: the dynamic behavior, the speech and the facial expression. The communicative power of faces implies that they are a central point of interest during social exchanges and become the most accessible window of the mechanisms that govern our emotional states. Some of the recent studies concerning HM interaction improvements focus on facial expression analysis to establish a real interactive dialogue between man and machine. In addition, the emergence of facial animation or virtual avatar creation imposes the necessity of robust representation, analysis and classification tools for facial expression recognition. From a physiological angle, a facial expression results from the facial feature deformation, due to the contraction and relaxing of facial muscles. The construction of a recognition tool depends on the significance given to facial expressions. A lot of disciplines have looked into this problem, such as philosophy, biology, psychology or psychoanalysis. Nevertheless, one of the most famous facial expression representation and

*Corresponding author.

E-mail addresses: sdubuiss@hds.utc.fr (S. Dubuisson), fdavoine@hds.utc.fr (F. Davoine), mmasson@hds.utc.fr (M. Masson).

0923-5965/02/$ - see front matter
PII: S0923-5965(02)00076-0

coding system, called facial action coding system (FACS), has been proposed in 1978 by Ekman and Friesen [10]. Their objective was to describe all the visually distinct local facial muscle movements using 46 actions units (AUs): each AU is associated with the physical behavior of a specific facial muscle, giving an appearance-based description of faces. In that way, a facial expression corresponds to the combination of a set of AUs. Ekman has also classified all the facial expressions into six prototypical universal categories: fear, surprise, anger, disgust, joy and sadness.

In this article, we focus on the prototypical classification of facial expressions, that includes three fundamental problems. Analysis of facial expressions requires a number of preprocessing steps which attempt to detect, or track the face, then locate and extract characteristic facial regions such as eyes, mouth and nose. Then, this information has to be represented, so that the different facial expressions are precisely described (edges, bunch graphs, motion vectors, etc.), to make the classification easier. Finally, the choice of classification tools, adapted to the representation, is a fundamental step for optimizing the facial expression recognition. The classification approaches are generally based on statistical or rule-based decision methods, depending on the representation chosen. Graph modeling provides a geometrical two-dimensional (2D) or three-dimensional (3D) description of faces, including topographic information: the principle is to position nodes on facial points, also called fiducial points, and to connect nodes with edges. Graph matching algorithms [12] can then measure a similarity between each node of a graph and the nodes of a general graph representation of expressions. The facial expression of an unknown face is recognized if its representation yields the highest similarity with the specific expression graph model. Motion analyzing methods focus on computing motion of either facial muscles or facial features between neutral and apex instances of a face. Optical flow estimation algorithms [29,31] compute dense motion fields in the entire face (or selected area): these motion vectors are mapped to facial expressions using motion templates which have been extracted by summing over a set of learned motion fields. Feature tracking [30] allows to estimate motion only over a selected set of prominent features. Finally, 2D or 3D models of the face can be adapted to the image data to estimate object motion [3,11]. Neural network approaches for facial expression recognition [21,18,26] use data collected during the past to construct a model (learning step), that is then applied to new test samples to estimate the posterior probability of each class of expression.

In this paper, we propose two kinds of contributions. We first show how to construct a discriminant representation subspace, adapted to a specific classification task. We illustrate its interest in the case of the three- and six-facial expression recognition problems. We also propose a decision tree classification, which is trained by an iterative selection of individual features that are more salient at each node of the tree. The organization of this paper is as follows. Section 2 briefly exposes some of the related work concerning the statistical facial expression recognition problem. In Section 3, we introduce the construction of a discriminant representation subspace that is well adapted to a given recognition task, and Section 4 focuses on the facial expression recognition problem: we first describe our data sets, composed of normalized internal parts of faces. We then study and compare the discriminant power of different representation subspaces. We present the comparison of classification performances, using an Euclidean distance-based classifier, into these representation subspaces. We also test the influence of different facial areas and different subspace sizes on the classification performances. Section 5 deals with the construction of a decision tree classifier, based on a ''coarse-to-fine'' classification approach and presents experimental results. Finally, concluding remarks are given in Section 6.

2. Facial expression recognition by statistical analysis

In a statistical approach, a face, or sample, is modeled as a d-dimensional feature vector. Statistical classification methods consider the statistical distribution of data into their original space or

into a low-dimensional subspace. Such analysis preserves the original images as much as possible in order to allow the classifier to discover the revealing features in the images. In general, statistical classification is divided into two steps: representation (learning) and recognition (testing).

Learning can consist in extracting or selecting features to find an appropriate representation of the input patterns, and then in training a classifier to partition the feature space. The dimensionality reduction is a well-known approach for data representation, whose goal is to decrease the measurement cost and increase the correct classification rate. One can however make a distinction between feature selection (in the measurement space) and feature extraction (equivalent to feature selection in a transformed space), even if both perform a dimensionality reduction [28]. Feature selection methods [13] consist in selecting the best subset of features among the input features: given a set of d features, they find a subset of size N that leads to the smallest classification error. Feature extraction methods (principal component analysis (PCA) [27,8], linear discriminant analysis (LDA) [2,15], independent component analysis (ICA) [1], local feature analysis (LFA) [19]) transform the original features into new ones, via linear combinations: they find an appropriate representation subspace, of dimension N, for the original d-dimensional patterns. Edwards et al. [9] have also interpreted face images using active appearance models (AAM), that represent both shape and gray-level appearances. The shapes of the main features and the spatial relationships between them are represented by a point distribution model, corresponding to a statistical model of shape variation computed by PCA. The statistical model of gray-level appearance is built by warping each training image, using triangulation, and then applying PCA to the shape-free images. By performing PCA once more on gray-level appearance and shape appearance models, a vector of appearance controlling both gray level and shape is generated, which provides a compact and parameterized description of a face. Statistical learning can also perform a density estimation of different classes to obtain a distribution model [17].

Once the data are modeled, the classification process consists in assigning a d-dimensional pattern into one of c classes $\{\omega_1, \ldots, \omega_c\}$: a classifier can be designed using three main approaches, which have been employed for facial expression recognition. The simplest is based on the concept of similarity [15,22,16]: samples that are similar should be assigned to the same class. This implies to use a good metric like the usual Euclidean or Mahalanobis distances. The second main approach is the probabilistic one [7] which requires the estimation of the parametric distribution models of the patterns. The Bayes decision rule, for the $\{0,1\}$ cost, assigns a sample to the class with the maximum posterior probability. For example, Moghaddam and Pentham [17] have used eigenspace decompositions to estimate the distribution of high-dimensional data as a mixture of Gaussian densities, and maximize a likelihood measure to recognize new patterns. The last concept consists in constructing decision boundaries directly by optimizing an error criterion: k-nearest neighbor rule-based techniques [24] and neural networks are well-known tools for such a process.

In this paper, we have adopted the statistical analysis approach: the next section describes the construction of a representation subspace well adapted to a specific class recognition problem.

3. Construction of an optimal subspace

We propose a method whose goal is to construct a projection subspace adapted to a specific recognition task. PCA and LDA play a critical role in many pattern classification tasks. Starting from the full signal space, and considering a learning set containing c different class samples, we first perform a dimensionality reduction by applying PCA. We then search for the most discriminant projection along eigenvectors by successively selecting the principal components, in the order of their importance for the recognition task. Finally, LDA is computed into this so-called *sorted eigenspace*, to generate a $(c-1)$-dimensional discriminant subspace where new samples are classified.

3.1. Principal component analysis (PCA)

One of the most successful approaches to the problem of creating a low-dimensional image representation is based on the Karhunen–Loeve expansion, well known as PCA. PCA is an unsupervised linear feature extraction method that generates a set of orthogonal basis vectors, which describe major variations in the whole training set, and where the mean square reconstruction error is minimized. Kirby and Sirovich have developed a technique using PCA to efficiently represent human faces [23]. Given a set of different face images, the technique first finds the principal components of the distribution of faces, expressed in terms of eigenvectors: each individual face can then be approximated by a linear combination of the largest eigenvectors, using appropriate weights. Turk and Pentland have later developed this technique for face recognition [27]. Since the face reconstruction is an approximation, the residual error so-called distance-from-feature-space, gives a good indication of face existence in an image.

Let $S = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N]$ be the centered learning set (the mean vector of the set has been subtracted to each of its vectors) containing N d-dimensional face vectors (e.g. the pixels of the face images in their lexicographical order) and let $C = SS^{\mathrm{T}}$ be its covariance matrix. PCA seeks the linear transformation matrix W_1 that maps the original space onto an N-dimensional subspace, with $N \ll d$, by factorizing the covariance matrix into the form $C = W_1 \Lambda W_1^{\mathrm{T}}$, where $W_1 = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_N]$ is the orthogonal nonzero eigenvector matrix of C and $\Lambda = \mathrm{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_N\}$ the corresponding diagonal eigenvalue matrix with diagonal elements sorted in decreasing order ($\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_N$).

The eigenvectors $\boldsymbol{v}_i$ ($i = 1, \ldots, N$), often referred to as *eigenfaces* in face analysis, are mutually orthogonal and span an N-dimensional subspace called *eigenspace*. By ordering the eigenvectors in the order of eigenvalues, one can create an orthogonal basis with eigenvectors having the largest variance direction in the learning set. In this way, we can find directions in which the learning set has the most significant amounts of energy. Nevertheless, instead of using all the eigenvectors, we may represent the data in terms of few basis vectors of W_1: a low-dimensional representation of the face images with minimum reconstruction error (MSE) is obtained by projecting the images onto the first few eigenvectors, according to the percentage of inertia they cover. The minimum number M of eigenvectors ($M < N$) is determined in order to keep the inertia ratio $r_M = (\lambda_1 + \lambda_2 + \cdots + \lambda_M)/\sum_{i=1}^{N} \lambda_i$ larger than a given threshold ζ. If we denote by W_{1_M} the matrix containing the subset of the first M eigenvectors, a d-dimensional centered input vector $\boldsymbol{x}$ can be linearly transformed into a M-dimensional vector $\boldsymbol{y} = W_{1_M}^{\mathrm{T}} \boldsymbol{x}$. This lower-dimensional vector $\boldsymbol{y}$ captures the most expressive features of the original data $\boldsymbol{x}$. The *principal components* correspond to the coordinates of the projected vectors onto the eigenspace: each of them grasps a type of information (illumination or feature position) concerning the original data set [20]. The next section describes the selection of principal components, to keep only the most discriminant ones.

3.2. Selection of the principal components

PCA is a feature extraction unsupervised method in the sense that it does not use any class information: eigenvectors with the largest eigenvalues are likely to convey information that is common to all samples, not to class categories. In a second step, following PCA, our objective is to select a feature combination that separates the class volumes so that the classes can be easily distinguished: maximizing the classification accuracy while minimizing the number of features. This selection process is commonly used in pattern recognition problems [28], but has not been proposed, according to our knowledge, for face or facial expression recognition. The method we propose for such a task is described in the next sections.

3.2.1. Principle of a forward stepwise selection

We consider here a training set of vectors, distributed into c classes. Each vector is then projected in an eigenspace (computed by PCA), spanned by N eigenvectors. The selection method

consists in seeking, among the N principal components, the K components which are most discriminant for the specific recognition problem, which are called "optimal". We use an iterative process that successively selects components step by step to construct an optimal sorted set: during step j ($j = 1, \ldots, N$), we seek the component, among the $(N - j + 1)$ available, which, when added to those previously selected, forms an optimal set of components.

The selection criterion F used to define the optimality of a set of components is a general class separability measure, defined by the Fisher criterion, which is expressed as $F = |S_B|/|S_W|$, where $|S_W|$ and $|S_B|$ are, respectively, the determinant of the within- and between-class scatter matrix. This criterion has to be maximized in order to select the best discriminant principal components. If $\boldsymbol{y}_i^k$ denotes the j-dimensional feature vector (e.g. principal component), extracted from the ith projected sample of the kth class c_k, composed of N_{c_k} samples, let $\boldsymbol{g}_k$ ($k = 1, \ldots, c$) be the mean vector of the kth class and $\boldsymbol{g}$ the total mean vector in this j-dimensional projection feature space, respectively, given by

$$\boldsymbol{g}_k = \frac{1}{N_{c_k}} \sum_i \boldsymbol{y}_i^k \quad \text{and} \quad \boldsymbol{g} = \frac{1}{c} \sum_{k=1}^{c} \boldsymbol{g}_k. \tag{1}$$

The within- and between-class scatter matrix can be calculated in this feature space as follows:

$$S_W = \sum_{k=1}^{c} \sum_{i=1}^{N_{c_k}} (\boldsymbol{y}_i^k - \boldsymbol{g}_k)^T (\boldsymbol{y}_i^k - \boldsymbol{g}_k)$$

and

$$S_B = \sum_{k=1}^{c} (\boldsymbol{g}_k - \boldsymbol{g})^T (\boldsymbol{g}_k - \boldsymbol{g}). \tag{2}$$

In order to avoid overfitting and to achieve better generalization performances, the selection criterion is computed as the average of the Fisher criterion F over several learning sets sampled from the original data set. Moreover, we estimate a generalization error rate in order to select, at the end, an optimum number of principal components. Note that we have chosen to use the Fisher criterion to select the optimal set of components. The classification error rate could have been used for such a selection, but the Fisher criterion seems to exhibit more stability (e.g. more "smoothness") than the classification error rate, especially when the size and the number of validation sets are small.

The measure of the classifier performance is the classification error rate e^j: the percentage of test samples that are assigned to the wrong class, based on the Mahalanobis distance between a sample

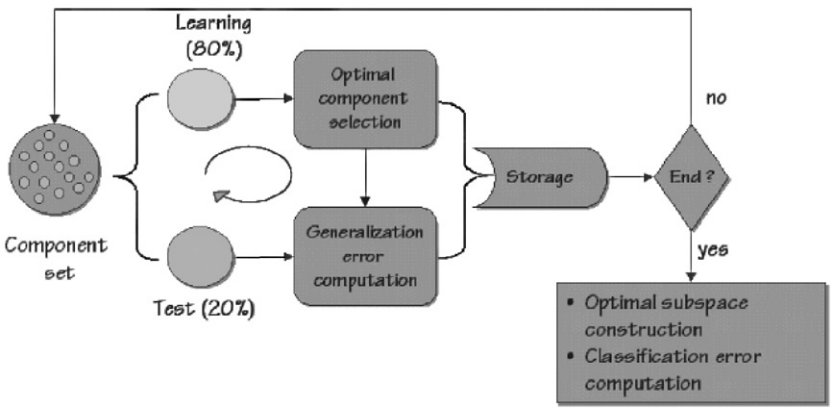

Fig. 1. General scheme of the sorting process: a randomly chosen learning set is used to determine, at each iteration, which of the available principal components is the optimal one. The complementary test set is then used to update a generalization classification error rate. This process is repeated until all the components have been sorted.

and its closest class, in the j-dimensional subspace (e.g. we have select j principal components, among the N available). The generalization error rate $\tilde{e}^j_{\text{gen}}$ in this subspace is the average over N_{iter} iterations (in our experiments, $N_{\text{iter}} = 40$) classification error rates: $\tilde{e}^j_{\text{gen}} = (1/N_{\text{iter}}) \sum_{k=1}^{N_{\text{iter}}} e^j_k$.

The global algorithm that selects all the components in decreasing order of their importance for a recognition problem is given below, and illustrated by Fig. 1.

For $j = 1$ to N components to select
  Set $F^j = 0$
  For iter $= 1$ to N_{iter} iterations for the selection of one component
    Randomly choose a test set (20% of the available samples set)
    Take the remaining 80% for the learning set
    For pc $= 1$ to $N - j + 1$ available principal components
      Add the tested principal component to those previously kept
      Build the eigenspace with the corresponding eigenvectors
      Project the learning set
      Compute the within-class scatter matrix S_{W}
      Compute the between-class scatter matrix S_{B}
      Compute the Fisher criterion value: $f^{\text{pc}}_{\text{iter}} = \frac{|S_{\text{B}}|}{|S_{\text{W}}|}$
      Project the validation set
      Compute the classification error rate $e^{\text{pc}}_{\text{iter}}$ into the eigenspace
      $F^{\text{pc}} = F^{\text{pc}} + f^{\text{pc}}_{\text{iter}}$
    End For
  End For
  Select the principal component pc* with maximum average Fisher criterion over the N_{iter} iterations: pc* $= \text{Argmax}_k \frac{F^k}{N_{\text{iter}}}$
    → Add pc* to those previously kept to form a j-dimensional subset
  Compute the generalization error rate $\tilde{e}^j_{\text{gen}}$ at level j
End For

The minimal number of features needed to construct the projection basis is then determined as follows: once all the principal components have been sorted, the final dimension K of the optimal subspace corresponds to the minimum $\tilde{e}^K_{\text{gen}}$ generalization error rate profile (see Fig. 2). We therefore seek the rank K for which the addition of a new component to the optimal set does not decrease the generalization error anymore: this provides an optimistically biased estimation of the error rate. Fig. 2 illustrates the case for different two-class problems (70 samples). The first graph ("Sadness/Joy" case) shows that we need $K = 17$ components to minimize the generalization error rate ($\tilde{e}^K_{\text{gen}} = 2\%$). Descriptions of the data set used for this experiment will be given in Section 4.

The optimal subspace is then constructed using the K corresponding eigenvectors, and from now, is called the *sorted eigenspace*. Fig. 3 shows two examples of two-dimensional projections of the learning set, along the first two axes of different projection subspace bases (eigenspace or sorted eigenspace). We note that sorting the principal components after PCA provides a representation where the two classes are better separated (second graph) than with the PCA representation alone (first graph). If we compute the Fisher criterion value for these two representation subspaces, we obtain $|S_{\text{B}}|/|S_{\text{W}}| = 1.17$ in the eigenspace, and $|S_{\text{B}}|/|S_{\text{W}}| = 3.31$ in the sorted eigenspace.

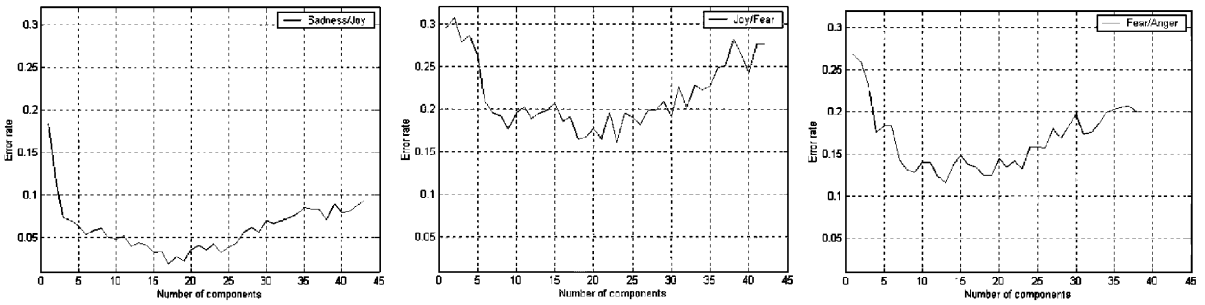

Fig. 2. Generalization error rate profiles for different two-class learning sets (70 samples: 35 face vectors by expression, and two facial expressions): finding its minimum allows to determine the optimal number K of needed components that provide the minimal classification error rate $\tilde{\epsilon}^K_{\text{gen}}$. From left to right, "Sadness/Joy" ($K = 17$, $\tilde{\epsilon}^K_{\text{gen}} = 2\%$), "Joy/Fear" ($K = 23$, $\tilde{\epsilon}^K_{\text{gen}} = 16\%$) and "Fear/Anger" ($K = 13$, $\tilde{\epsilon}^K_{\text{gen}} = 12\%$).

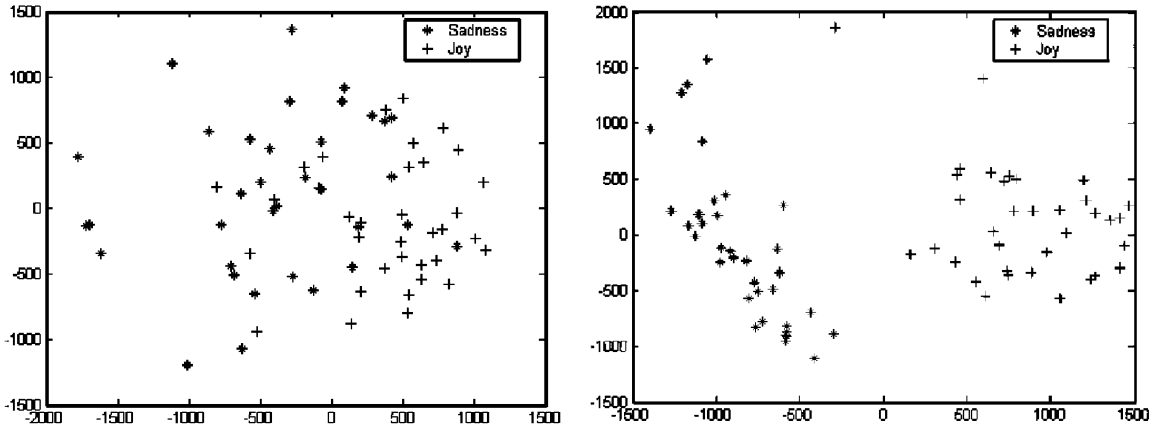

Fig. 3. The 2D-projections of the two-class learning set "Sadness/Joy" (70 samples) onto the subspaces which basis corresponds to, from left to right, the first two eigenvectors and the two optimal eigenvectors.

Fig. 4. From top to bottom, the first eight eigenfaces and the first eight optimal ones (after the sorting process). The learning set contains $N = 210$ facial masks: $N_c = 35$ per facial expression for six expressions.

Consequently, the first two optimal principal components make a better use of class information than the two components associated with the largest eigenvalues. Fig. 4 shows, on its first row, the first eight eigenfaces generated by PCA applied to a learning set containing six facial expression classes ($N = 210$ samples). The second row shows the first eight optimal eigenfaces, after the selection

process: they correspond, from left to right, to the PCA eigenfaces (sorted in decreasing order of their eigenvalues) number 2, 1, 3, 11, 4, 14, 5 and 8. One can see that the six universal facial expressions are visible in the first six optimal eigenfaces, from left to right: anger, fear, surprise, disgust, joy and sadness. These facial expressions do not necessarily appear in the first six eigenfaces generated by PCA.

We have just described a method allowing to construct a discriminant subspace, adapted to a specific recognition task.

The next step of the method, described in the following section, consists in performing LDA into the K-dimensional sorted eigenspace.

3.3. Linear discriminant analysis (LDA)

Face recognition systems using LDA have been very successful [25,2,4], where training is carried out via scatter matrix analysis. The prime difference between LDA and PCA is that in PCA, the shape and location of the original data set change when transformed to a different space, whereas LDA does not change the location but only tries to provide more class separability. LDA searches for those vectors in the underlying space that best discriminate among classes: given a fixed number of features describing the data, LDA looks for the linear combination maximizing some measure J of class separation or classification performance. Various measures J are available for quantifying the discriminatory power, a commonly used one being the ratio of the determinant of the between-class scatter matrix to the within-class scatter matrix: $J(W_2) = |W_2^{\mathrm{T}} S_{\mathrm{B}} W_2|/|W_2^{\mathrm{T}} S_{\mathrm{W}} W_2|$, where S_{W} and S_{B} are the within- and between-class scatter matrix (see Eq. (2)), respectively, and W_2 denotes the optimal projection matrix. W_2, which maximizes the ratio, can be derived by solving the generalized eigenvalue problem: $S_{\mathrm{B}} W_2 = \Lambda S_{\mathrm{W}} W_2$, where W_2 contains the eigenvectors of $S_{\mathrm{W}}^{-1} S_{\mathrm{B}}$, and Λ is the diagonal eigenvalue matrix. For a c-class problem, column vectors of W_2, well known as *Fisherfaces*, form the basis of the optimal $(c-1)$-dimensional discriminant subspace: each of them has captured discriminant information contained in the learning set.

It should be noted that if the number of samples is too small, compared to the dimensionality of the samples, S_{W}^{-1} is very close to being singular, and consequently, LDA should not be applied directly to the input samples. That is the reason why, as suggested by Belhumeur et al. [2], most discrimination methods [15,16] use PCA subspace projection as a first step in processing the face data. Projecting an M-dimensional face vector $\boldsymbol{y}$ (see Section 3.1) into the LDA subspace yields a $(c-1)$-dimensional vector $\boldsymbol{z}$ such that $\boldsymbol{z} = W_2^{\mathrm{T}}\, \boldsymbol{y}$. Vectors represented in this subspace are then directly classified.

In the same way, our approach consists in first performing a dimensionality reduction by sorting the principal components after PCA, and keeping the K most discriminant, and secondly in applying LDA into the sorted eigenspace. The $(c-1)$-dimensional generated subspace will be called *sorted eigenspace plus Fisherspace* (SE + F), and the process is, from now, called *Sorted PCA plus LDA method*. Fig. 5 shows the five Fisherfaces generated by Sorted PCA plus LDA method, applied to a learning set containing six facial expression classes (210 samples). It can be seen that the discriminant information, concerning

Fig. 5. Illustration of the five Fisherfaces, corresponding to five axes of the subspace generated by Sorted PCA plus LDA method, which are used as basis of the final discriminant subspace. The learning set contains $N = 210$ facial masks: $N_c = 35$ per facial expression for six expressions.

facial expressions, seems to be essentially concentrated on the bottom part of a face, around the mouth. On the other hand, the eyes do not seem to convey a lot of facial expression information.

4. Application: classification of facial expression

4.1. Data extraction

The proposed algorithm performs a statistical analysis via PCA and LDA: both techniques require precise normalization and registration of faces. Most of the facial expression information is concentrated around facial features such as the eyes or the mouth. Including irrelevant parts (hair, background,…) can generate incorrect decisions for expression recognition [5]. When we perform a statistical analysis to solve a recognition problem, we must consider normalized data: the classification mechanism may not depend on physiognomic variability of the observed persons: the variations between the samples must ideally be only due to the pattern we have to recognize. In practice, we have to minimize the other possible variations (feature positions, variations in illumination, etc.). That is the reason why we have aligned our faces, by performing a manual facial mask extraction: four facial feature points (pupil centers, top of nose and middle of mouth) have been chosen as relevant for normalization. Two affine transformations T_1 and T_2, applied independently, respectively, on the top and bottom parts of faces, are used in such a way that these four points are located in fixed positions in target images. T_1 transforms the eyes–nose triangle to a target triangle, and T_2 transforms the eyes–mouth triangle to a target triangle. We then crop the right and left lateral parts of faces to only consider their internal zone, of size 60×70, corresponding to 4200-pixel vectors. At last, we perform a histogram specification, using the histogram of a learned face as reference, to compensate for the variations in illumination and skin colors. Some examples of manually extracted facial masks are given in Fig. 6.

We use facial masks extracted from the CMU-Pittsburgh image database [14] to construct a learning set containing $N_c = 25 \times 2$ samples (facial mask and their mirror) per expression and six high intensity facial expressions: surprise, sadness, disgust, joy, fear and anger. We then have constructed different test sets using manually extracted facial masks from different databases (Fig. 6):

- *Test set 1*. Contains 194 people which do not belong to the learning set (CMU-Pittsburgh database) with high intensity facial expression.
- *Test set 2*. Contains 151 people from different databases (Yale, JAFFE [15], etc.), except the CMU-Pittsburgh one. The faces can have glasses, bears or mustaches, and high facial expression intensity.

4.2. Facial expression representation

4.2.1. Three-class problem

We built a learning set using $N = 105$ facial masks (Section 4.1) and consider a three-facial expression class problem (sadness, fear and disgust) with $N_c = 35$ 4200-dimensional vectors per class. Images from Fig. 7 show the 2D-projections of the learning set, along the first two axes of the basis, onto different subspaces: eigenspace (E), eigenspace plus Fisherspace (E + F), sorted eigenspace (SE) and sorted eigenspace plus Fisherspace (SE + F). We can visually see that the representation subspace that exhibits the best separation between the three classes is (SE + F).

Fig. 6. Extracted facial masks, for different facial expressions (templates of size 60×70: 4200-pixel vectors), from left to right, CMU-Pittsburgh and Yale databases.

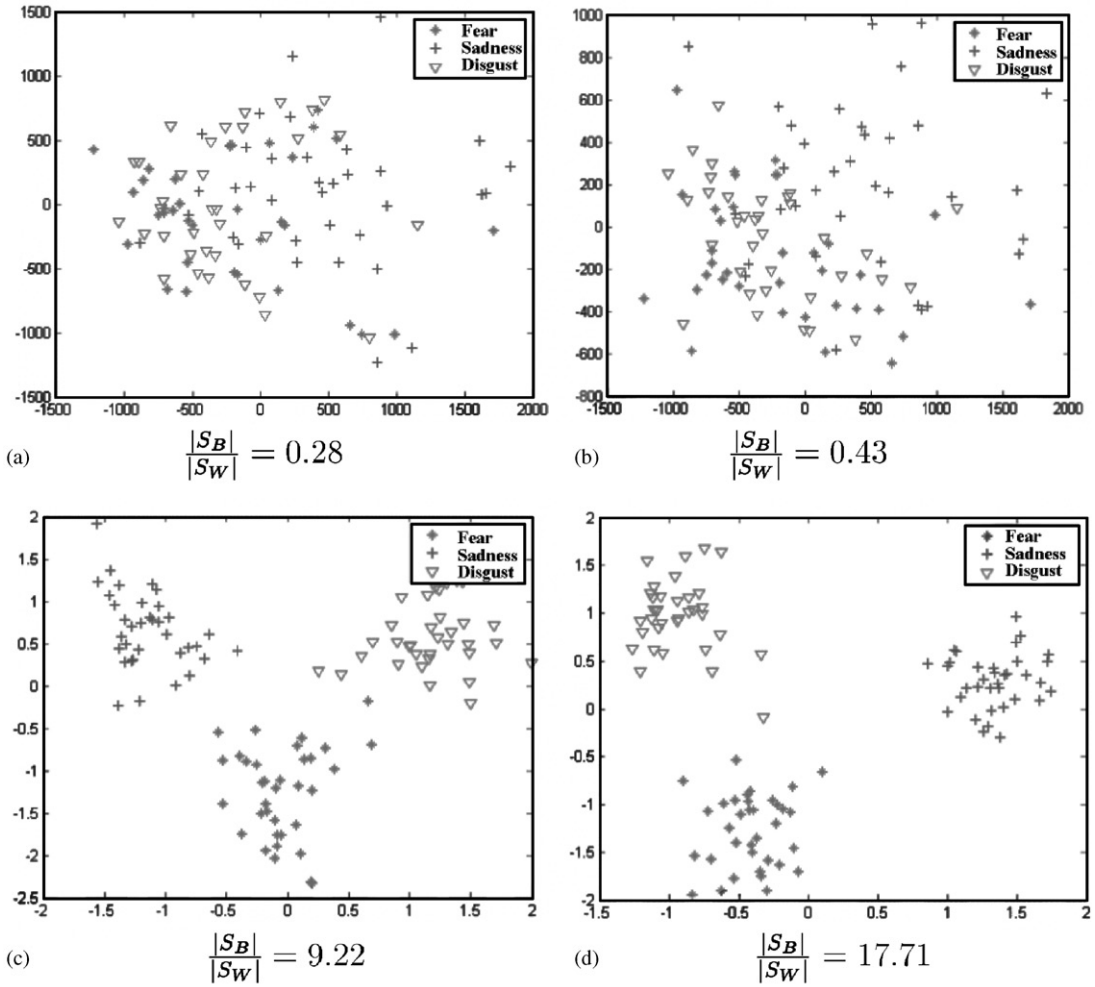

Fig. 7. 2D-projections of the three-class learning set "Fear/Sadness/Disgust" (105 samples) along the two first axes of the basis of: (a) the eigenspace (E); (b) the sorted eigenspace (SE); (c) the eigenspace plus Fisherspace (E + F); and (d) the sorted eigenspace plus Fisherspace (SE + F). Fisher criterion values computed in each of the five-dimensional subspaces are reported below each graph.

This can be confirmed by computing the Fisher criterion value in these different five-dimensional subspaces (see corresponding values below images in Fig. 7): the highest value is also obtained for (SE + F). In general, we can draw two conclusions concerning the effects of the sorting process:

(1) It improves the separation of the classes after PCA: the Fisher criterion value computed in the representation subspaces is generally twice as large for the sorted PCA compared to the regular PCA.

(2) It improves the discriminant power of LDA (Fig. 7).

Consequently, Sorted PCA method can be seen as a preprocessing step useful to improve discriminant capacities of both PCA and LDA.

4.2.2. Six-class problem

We built a learning set using $N = 300$ facial masks (Section 4.1) with $N_c = 25 \times 2$ 4200-dimensional vectors per facial expression class and six facial expressions. We first ran PCA on this data

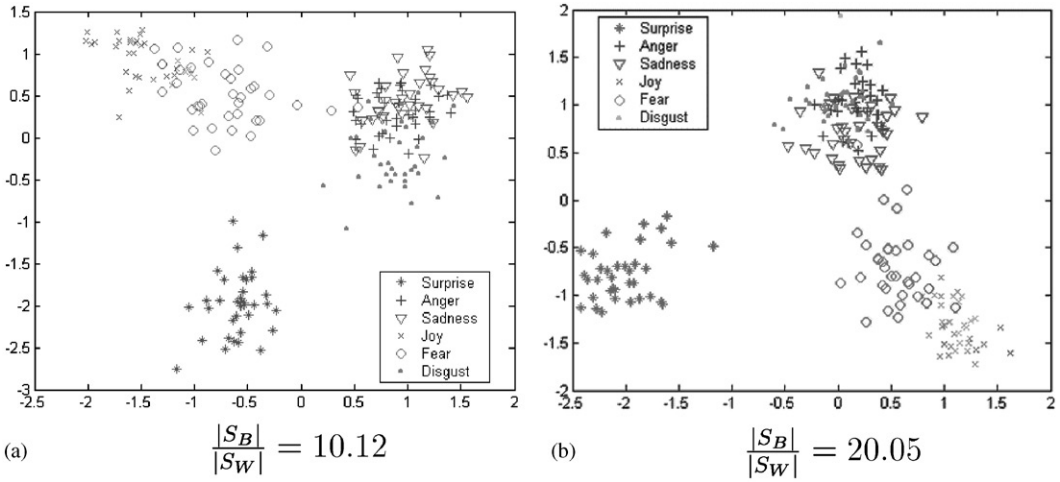

Fig. 8. 2D-projections of the six-facial expression learning set (300 samples) along the first two axes of the basis onto: (a) (E + F) and (b) (SE + F) subspaces. Fisher criterion values computed in the five-dimensional subspaces are reported below each graph.

set, and generate $N = 300$ principal components. The selection process then provided an optimal number $K = 58$ of needed components, sorted in decreasing order of their importance for this six-class problem: the projection from the image space to the sorted eigenspace maps from $\mathbb{R}^{4200}$ to $\mathbb{R}^{58}$. Then, the projection from sorted eigenspace to Fisherspace, by performing LDA, maps from $\mathbb{R}^{58}$ to $\mathbb{R}^5$: this discriminant subspace (SE + F) is used both for representation and classification. Fig. 8 shows the six-class learning set projected onto the two first components of the PCA plus LDA (E + F) and Sorted PCA plus LDA (SE + F) subspaces: we note that Sorted PCA plus LDA method doubles the Fisher criterion value compared to PCA plus LDA method.

4.3. Facial expression classification

4.3.1. Facial mask classification

The proposed method has been tested with the different test sets given in Section 4.1: 345 new facial masks (which do not belong to the learning set) from the CMU-Pittsburgh [14], Yale,[1] JAFFE [15] and other databases. The classifier recognizes the facial expression class of these new samples by using a measure of similarity between them and the different facial expression class centroids. After geometric and illumination normalizations (Section 4.1), we project the test samples onto the five-dimensional discriminant subspace (previous sections). We then compute the Euclidean distance to determine to which of the six facial expression classes they belong. Table 1 shows the classification performances, depending on the test set, into the five-dimensional discriminant subspace. We can see the performances are lower with test set 2, because faces do not belong to the same database than learning samples, and, in addition, people can have glasses, beards or mustaches. In Table 2, we present the comparative classification performances of the 345 samples belonging to test sets 1 and 2, for four different subspaces. We can see that the best performances are achieved with the (SE + F) subspace, where the correct classification rate is up to 85.5%: the contribution of our selection process increases the classification accuracy by 3% compared to "PCA plus LDA" method.

4.3.2. Forward versus backward selection of the principal components

Previous sections describe the forward selection of the principal components: starting with an

[1] http://cvc.yale.edu/projects/yalefaces/yalefaces.html.

Table 1
Correct classification performances, for the six universal facial expressions, into the five-dimensional discriminant subspace, depending on the test set (see Section 4.1). The classifier is based on the Euclidean distance measure

Expression	Surprise	Anger	Sadness	Joy	Fear	Disgust	Total
Test set 1 (194 samples)	96%	82%	94%	85%	81%	88%	87.6%
Test set 2 (151 samples)	85%	90%	80%	85%	72%	90%	83.6%
Mean (345 samples)	91%	86%	87%	85%	77%	89%	85.8%

Table 2
Correct classification performances (test sets 1 + 2), for the six universal facial expressions, into different representation subspaces. The classifier is based on the Euclidean distance measure

Expression	Surprise	Anger	Sadness	Joy	Fear	Disgust	Total
No. of test samples	88	27	55	109	38	28	345
(E)	88%	67%	75%	66%	63%	75%	72.3%
(SE)	88%	70%	78%	70%	66%	75%	74.5%
(E + F)	88%	86%	84%	80%	73%	84%	82.5%
(SE + F)	91%	86%	87%	85%	77%	89%	85.8%

Table 3
The 18 first optimal principal components, sorted in the order of their importance for the problem of six-facial expression recognition, depending on the selection process (forward or backward selection)

Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Forward	3	2	14	4	5	12	8	54	19	28	13	16	18	66	112
Backward	125	3	2	5	4	14	12	8	19	54	13	28	19	112	26

empty set, we progressively select the components according to their relevance for a recognition problem. Another well-known selection process consists in starting with the full set of principal components, and, step by step, remove the less discriminant ones. We have compared the classification results according to these two selection strategies into (SE + F) subspace. The characteristics of the sorted eigenspace are given below:

- *Forward selection*. The final dimension of the optimal subspace is $K = 58$, giving a generalization error rate of $\tilde{e}_{\text{gen}}^{K} = 12.22$.
- *Backward selection*. The final dimension of the optimal subspace is $K = 64$, giving a generalization error rate of $\tilde{e}_{\text{gen}}^{K} = 12.35$.

Table 3 shows the 18 first sorted principal components, according to the selection process. Their number corresponds to their rank when they are sorted in decreasing order of their corresponding eigenvalue. The Fisher criterion value, computed in the five-dimensional discriminant subspace is equal to 20.05 after a forward selection and is equal to 21.11 after a backward selection. We can see that the two ways of selecting the principal components generate comparable discriminant subspaces. However, Table 4 shows that these two subspaces do not characterize the same facial expression classes. For example, the classification performances are reversed for the "Anger" and "Sadness" classes.

Table 4
Correct classification performances, for the six universal facial expressions, into the five-dimensional (SE + F) subspace, depending on the selection process

	Surprise	Anger	Sadness	Joy	Fear	Disgust	Average
Forward	91%	86%	87%	85%	77%	89%	85.8%
Backward	90%	91%	84%	84%	80%	85%	85.6%

Other hybrid stepwise selection strategies could have been used, considering both forward and backward moves at each stage and making the best move (addition or suppression of principal components). A time consuming branch-and-bound procedure could have been used to find a subset which is guaranteed to be the best. These strategies are all heuristics to avoid considering all possible optimal subsets of principal components.

4.3.3. Influence of the size of the eigenspace

We also have compared the evolution of the classification error rate, for the test set, depending on the dimension of the PCA and Sorted PCA subspaces. Fig. 9 shows the results of this comparative study. As expected, the average recognition error rate decreases with the number of eigenfaces used in the projection for both subspaces. Nevertheless, we note substantial behavioral differences of the classification error evolution for the two subspaces:

- *Eigenspace* (*PCA*). We can see that the classification error rate rapidly decreases to a minimum of $e = 0.23$ for a 20-dimensional eigenspace. Then it stabilizes.
- *Sorted eigenspace* (*Sorted PCA method*). The classification error rate decreases much more slowly to a minimum of $e = 0.21$ for a 50-dimensional sorted eigenspace. This error then increases more quickly.

We observe surprisingly on this figure two points for which the PCA-based classification error rate is better than the sorted PCA-based one. This could be explained by the fact that the learning set (used for the selection) and the test set (used in this experiment) have slightly different distributions. However, note that the classification error rate for the eigenspace is 2% higher than for the sorted eigenspace: the sorting process provide a better classification rate. This experiment shows that the method we propose in this article actually provide an optimal dimension of a projecting subspace, beyond which the classification error increases.

4.3.4. Influence of different facial parts

We now present comparative recognition tests to determine which part of a face seems to be the most discriminant for the facial expression characterization. We have tested four different facial feature areas: the eyes (templates of size 30×60, 1800-pixel vectors), the mouth (templates of size 20×30, 600-pixel vectors), the eyes and the mouth (union of the two previous templates,

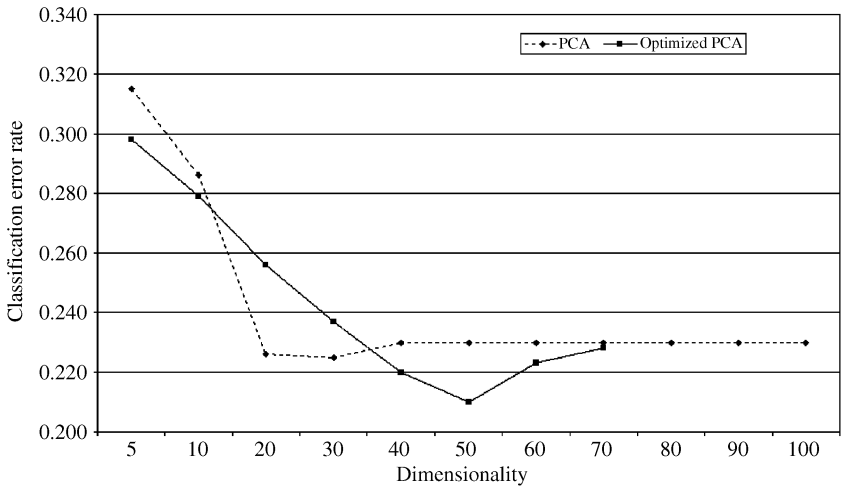

Fig. 9. Illustration of the influence of the dimension of the projection subspace on the classification error rate, depending on the learning method (PCA or Sorted PCA).

Fig. 10. Two of the tested facial feature area, from top to bottom: eyes (1800-dimensional vectors) and mouths (600-dimensional vectors).

Table 5
Influence of different facial parts on facial expression recognition rate. Four cases have been tested: eyes, mouth, eyes + mouth and facial masks

Expression	Surprise	Anger	Sadness	Joy	Fear	Disgust	Tot.
No. of test samples	88	27	55	109	38	28	345
Eyes	85%	63%	66%	70%	46%	59%	64.8%
Mouth	90%	78%	82%	75%	72%	82%	79.8%
Eyes + mouth	90%	88%	82%	75%	73%	82%	81.6%
Facial mask	91%	86%	87%	85%	77%	89%	85.8%

2400-pixel vectors) and the facial mask (4200-pixel vectors, Section 4.1). Fig. 10 gives examples of eyes and mouth regions for the six facial universal expressions. For each of these facial parts, we independently construct their representation and recognition subspace by applying Sorted PCA plus LDA method on a learning set containing $N = 300$ samples. Tests are performed using new samples of corresponding facial areas. Table 5 gives the facial expression classification performances, depending on the considered facial parts. We note that, as part of these tests, the entire internal part of a face (facial mask) seems to be more discriminant than other parts. These results corroborate the intuitive observations we made about Fisherfaces of Fig. 5 (Section 3.3). In fact, facial masks include transient facial features, such as wrinkles, that can appear on the regions surrounding the mouth (cheek): these transient features characterize some facial expressions particularly well, such as joy, anger and surprise. The eyes may not give sufficient information to provide a good description of the facial expression. According to these results, we have chosen to work, with the entire internal part of faces.

5. Optimization of the classification process

The previous sections have detailed the principle of the construction of a discriminant subspace adapted to a specific recognition problem. We have illustrated its interest in the case of the three- and six-facial expression recognition problems. In this section, we propose a classification process, using a decision tree classifier [6], that takes into account the properties of our representation subspace. This classifier is trained by an iterative selection of individual features that are more salient at each node of the tree. The fundamental problem when constructing a decision tree is to determine tree partitions based on the training data. The next section describes the partitions that we have chosen.

5.1. Discriminant subspace characteristics

If we look at the confusion matrix (Table 6) and measure the Mahalanobis distance between classes (Table 7) in the five-dimensional discriminant subspace (Sorted PCA plus LDA, see Section 3), we observe that some classes are very close to each other. Experimentally, we observe that the six

Table 6
Confusion matrix for the six-facial expression recognition problem

	Surprise	Anger	Sadness	Joy	Fear	Disgust
Surprise	80	0	2	0	0	6
Anger	0	23	0	0	0	4
Sadness	0	3	49	2	0	0
Joy	3	3	0	93	10	0
Fear	0	0	0	6	30	2
Disgust	0	1	2	0	0	25

Table 7
Mahalanobis distances between the six facial expression classes

	Surprise	Anger	Sadness	Joy	Fear	Disgust
Surprise	0	253	118	403	108	117
Anger	—	0	12	175	153	22
Sadness	—	—	0	122	129	50
Joy	—	—	—	0	18	131
Fear	—	—	—	—	0	74
Disgust	—	—	—	—	—	0

facial expression classes can be regrouped into three main clusters: Group1 (G_1: Surprise), Group2 (G_2: Anger, Sadness and Disgust), Group3 (G_3: Joy and Fear).

According to these properties, we propose to construct a decision tree classifier [6], whose principle is explained in the next section.

5.2. Decision tree classifier principle and construction

We consider three representation subspaces, all of them are generated using the Sorted PCA plus LDA approach, described in Section 3.

- The "coarse" representation subspace S_G. It gives a representation of the three class groups (2D subspace): Group1 (Surprise), Group2 (Anger, Sadness and Disgust) and Group3 (Fear and Joy). These class groups, respectively, contain $N_{c_1} = 50$, $N_{c_2} = 150$ and $N_{c_3} = 100$ training samples.
- The representation subspace S_{G_2} (2D), that gives a finer representation of the three classes belonging to Group2. It is constructed with $N = 3 \times 50 = 150$ training samples.
- The representation subspace S_{G_3} (1D), that gives a finer representation of the two classes belonging to Group3. It is constructed using $N = 2 \times 50 = 100$ training samples.

The tree classification consists in first projecting a new sample onto the "coarse" representation subspace S_G and in associating it with the nearest class group, using a Euclidean distance-based classifier. Secondly, the sample is projected onto the subspace S_{G_i}, $i = \{2, 3\}$, representing the classes belonging to the group. Again, the Euclidean distance is used to determine the correct facial expression class. The different representation subspaces are shown in Fig. 11: a new sample is first projected onto S_G (middle graph of Fig. 11). If its projection is closer to Group1, it is classified as Surprise. If it is closer to Group2, it is projected onto S_{G_2} (left graph of Fig. 11), then classified into the nearest facial expression class (Fear, Sadness or Disgust). If it is closer to Group3, it is projected onto S_{G_3} (right graph of Fig. 11), then classified into the nearest facial expression class (Joy or Fear). The pseudo-algorithm applied to classify a new sample $\boldsymbol{x}$ is described below:

Define $\boldsymbol{y}$ as the projection of $\boldsymbol{x}$ onto S_G
If $\boldsymbol{y} \in G_1$
  class $\leftarrow$ Surprise
Else If $\boldsymbol{y} \in G_2$
  Define $\boldsymbol{z}$ as the projection of $\boldsymbol{y}$ onto S_{G_2}
  Classify $\boldsymbol{z}$ into the nearest class
Else
Define $\boldsymbol{z}$ as the projection of $\boldsymbol{y}$ onto S_{G_3}
  Classify $\boldsymbol{z}$ into the nearest class
End if

5.3. Results

We have applied the tree classification process, described in Section 5, on the six-facial expression recognition problem. We then have compared two classifiers: PCA plus LDA Euclidean distance-based classifier (where all six classes are considered at once) and tree classifier. The results are given in Table 8: the decision tree classifier yields a correct classification rate that is 5% larger than Euclidean

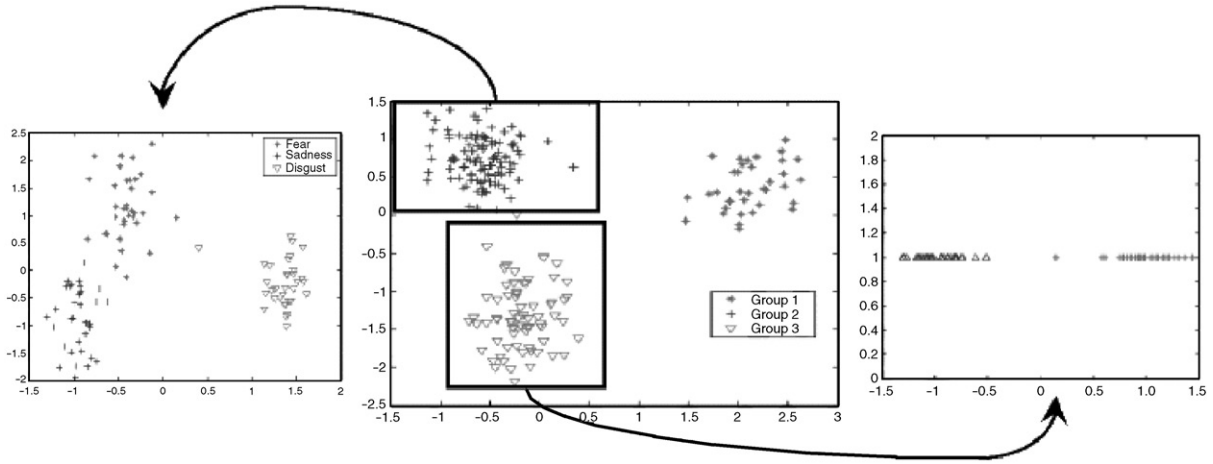

Fig. 11. The tree classification is carried out by successive projections: a new sample is first projected onto a "coarse" representation subspace, where it is associated with the closest class group of facial expression. The sample is then projected onto a finer representation subspace, describing the classes belonging to the group, to recognize its expression. The classification is based on the Euclidean distance.

Table 8
Classification performance, for manually extracted facial masks (4200-pixel vectors), depending on the classifier: Euclidean distance into (SE + F) subspace, or decision tree classifier

Expression	Surprise	Anger	Sadness	Joy	Fear	Disgust	Tot.
No. of test samples	88	27	55	109	38	28	345
(E + F)	88%	86%	84%	80%	73%	84%	82.5%
Tree classification	90%	91%	89%	88%	83%	85%	87.6%

distance classifier. This tree classification process is particularly efficient if classes overlap.

6. Conclusion

In this paper, we have proposed a statistical-based technique whose goal is to construct a discriminant representation subspace adapted to a specific recognition problem. We have illustrated its performance for the problem of facial expression recognition: this study shows that choosing an optimal representation for faces within the principal component approach can improve the recognition task. Tests have proven quantitatively and qualitatively the interest in sorting the principal components, in the order of their importance for a recognition task, before applying LDA. We have then proposed a decision tree process that provides a "coarse to fine" classification, increasing the classification accuracy by 5% for the six-facial expression recognition problem.

Preliminary tests have been performed using a fully automatic facial feature detection algorithm, in order to normalize vectors before classification. The tests reveal that such input face vectors are less separated in the feature space: the classification error rate is 10% larger. It seems that, for our actual system, the normalization phase of the samples is important to achieve good classification results. We then stress that the good performances of the recognition method strongly depend on the precision of the facial feature extraction step, that we did here manually. However, we are currently investigating the effect of the proposed representation to improve correct classification rates.

References

[1] M. Bartlett, H. Lades, T. Sejnowski, Independent component representations for face recognition, in: Proceedings of the SPIE, Conference on Human Vision and Electronic Imaging III, 1998, Vol 3299, pp. 528–539.

[2] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 711–720.

[3] M. Black, Y. Yacoob, Recognizing facial expressions in image sequences using local parameterized models of image motion, Internat. J. Comput. Vision 25 (1) (1997) 23–48.

[4] K.E.R. Chellappa, Discriminant analysis for recognition of human face images, J. Opt. Soc. Am. A (1997) 1724–1733.

[5] L.-F. Chen, H.-Y.M. Liao, M.-T. Ko, J.-C. Lin, C.-C. Han, Why recognition in a statistical-based face recognition system should be based on the pure face portion: a probabilistic decision-based proof, Pattern Recognition 33 (2000) 1713–1726.

[6] P. Chou, Optimal partitioning for classification and regression trees, IEEE Trans. Pattern Anal. Mach. Intell. 13 (4) (1991) 340–354.

[7] A. Colmenarez, B. Frey, T. Huang, Detection and tracking of faces and facial features, in: International Conference on Image Processing, Kobe, Japan, 1999.

[8] G. Cottrell, J. Metcalfe, Empath: face, gender and emotion recognition using holons, Adv. Neural Inform. Process. Systems 3 (1991) 564–571.

[9] G. Edwards, T. Cootes, C. Taylor, Face recognition using active appearance models, in: Proceedings of European Conference on Computer Vison, 1998, Vol. 2, pp. 581–595.

[10] P. Ekman, W. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movements, Consulting Psychologists Press, California, 1978.

[11] I.A. Essa, A.P. Pentland, Coding, analysis, interpretation and recognition of facial expressions, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 757–763.

[12] H. Hong, H. Neven, C. von der Malsburg, Online facial expression recognition based on personalized galleries, in: Proceedings of the Third International Conference of Face and Gesture Recognition, Nara, Japan, 1998, pp. 354–359.

[13] A. Jain, D. Zongker, Feature selection: evaluation, application and small sample performance, IEEE Trans. Pattern Anal. Mach. Intell. 19 (2) (1997) 153–158.

[14] T. Kanade, J.F. Cohn, Y. Tian, Comprehensive database for facial expression analysis, in: Proceedings of the Fourth International Conference of Face and Gesture Recognition, Grenoble, France, 2000, pp. 46–53.

[15] M. Lyons, J. Budynek, S. Akamatsu, Automatic classification of single facial images, IEEE Trans. Pattern Anal. Mach. Intell. 21 (12) (1999) 1357–1362.

[16] A. Martínez, Semantic access of frontal face images: the expression-invariant problem, in: Proceedings of IEEE Workshop on Content-Based Access of Images and Video Libraries, 2000, pp. 55–59.

[17] B. Moghaddam, A. Pentland, Probabilistic visual learning for object recognition, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 696–709.

[18] C. Padgett, G. Cottrell, Identifying emotion in static face images, in: Proceeding of the 2nd Joint Symposium on Neural Computation, University of California, San Diego, 1995, Vol. 5, pp. 91–101.

[19] P.S. Penev, J.J. Atick, Local feature analysis: a general statistical theory for object representation, Network: Comput. Neural Systems 7 (3) (1996) 477–500.

[20] P. Penev, L. Sirovich, The global dimensionality of face space, in: International Conference on Automatic Face and Gesture Recognition, Grenoble, France, 2000.

[21] M. Rosenblum, Y. Yacoob, L. Davis, Human emotion recognition from motion using a radial basis function network architecture, in: IEEE Workshop on Motion of Non-Rigid and Articuled Objects, Austin, Texas, 1994.

[22] Y. Shinza, Y. Saito, Y. Kenmochi, K. Kotani, Facial expression analysis by integrating information of feature-point positions and gray levels of facial images, in: IEEE International Conference on Image Processing, Vancouver, Canada, 2000.

[23] L. Sirovich, M. Kirby, Low-dimensional procedure for the characterization of human faces, J. Opt. Soc. Am. 4 (3) (1987) 519–524.

[24] K.-K. Sung, T. Poggio, Example-based learning for view-based human face detection, IEEE: Trans. Pattern Anal. Mach. Intell. 20 (1998) 39–51.

[25] D. Swets, J. Weng, Using discriminant eigenfeatures for image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 7 (1996) 831–836.

[26] Y. Tian, T. Kanade, J. Cohn, Recognizing action units for facial expression analysis, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2) (2001) 97–115.

[27] M. Turk, A. Pentland, Eigenfaces for recognition, J. Cognitive Neurosci. 3 (1) (1991) 71–86.

[28] A. Webb, Statistical Pattern Recognition, Arnold Press, London, and Oxford Univ. Press, 1999.

[29] Y. Wu, T. Kanade, J. Cohn, C. Li, Optical flow estimation using wavelet motion model, in: Proceedings of the International Conference of Computer Visions, Bombay, India, 1998.

[30] Y. Yacoob, L. Davis, Recognizing human facial expressions, in: Proceedings of the 2nd Workshop on Visual Form, Capri, Italy, 1994, pp. 584–593.

[31] Y. Yacoob, L. Davis, Recognizing human facial expressions from long image sequences using optical flow, IEEE Trans. Pattern Anal. Mach. Intell. 18 (6) (1996) 636–642.