Recognizing Imprecisely Localized, Partially Occluded, and Expression Variant Faces from a Single Sample per Class

Aleix M. Martínez, *Member*, *IEEE*

Abstract—The classical way of attempting to solve the face (or object) recognition problem is by using large and representative data sets. In many applications, though, only one sample per class is available to the system. In this contribution, we describe a probabilistic approach that is able to compensate for imprecisely localized, partially occluded, and expression-variant faces even when only one single training sample per class is available to the system. To solve the localization problem, we find the subspace (within the feature space, e.g., eigenspace) that represents this error for each of the training images. To resolve the occlusion problem, each face is divided into k local regions which are analyzed in isolation. In contrast with other approaches where a simple voting space is used, we present a probabilistic method that analyzes how "good" a local match is. To make the recognition system less sensitive to the differences between the facial expression displayed on the training and the testing images, we weight the results obtained on each local area on the basis of how much of this local area is affected by the expression displayed on the current test image.

Index Terms—Face recognition, occlusions, expression-variant faces, learning from undersampled distributions, facial asymmetry, localization of faces, principal components analysis, pattern recognition.

1 Introduction

The importance of research on face recognition (FR) is fueled by both its scientific challenges and its potential applications. However, most of the systems designed to date can only successfully recognize faces when images are obtained under constrained conditions. Therefore, the general problem of FR remains to be solved. Some important contributions have been made recently that allow systems to compensate for some of the difficulties of an unconstrained environment, for example, to compensate for the illumination changes [3], [4], [12], [50] or the (3D) depth rotations [5], [43], [66].

This paper proposes several approaches to deal with some of the difficulties that one encounters when trying to recognize frontal faces in unconstrained domains and when only one sample per class is available to the learning system. This we will do within the appearance-base paradigm (i.e., where only the texture of the face image is considered) using the principal components analysis (PCA or eigenspaces or eigenfaces) approach [58], [28], [63], [43], [3]. The problems we tackle in this contribution are:

- *Imprecise localization of faces*. It is clear that every FR system requires a first localization stage. When an appearance-based method is applied we also require a warping stage to guarantee that all images can be interpreted as a vector [5], [35] (this is discussed in Section 2.3). However, all localization algorithms have an associated error, namely, they cannot localize every single face feature with pixel precision. Here, *localization error* refers to those *small* localization errors that make the feature vector (e.g., eigen-representation) of a test image close to an incorrect class. This effect is depicted in Fig. 1a. In this figure, we display two classes, one drawn using crosses and the other with pentagons. For each class there are two learned feature vectors, each corresponding to the same image but accounting for different localization errors. The test image (which belongs to the "cross" class) is shown as a square. Note that, while one of the "pentagon" samples is far from the test feature vector, the other corresponds to the closest sample, that is, while one localization leads to a correct classification, the other does not. This point becomes critical when the learning and testing images differ on facial expression, illumination conditions, etc., as well as for duplicates. A "duplicate" is an image of a face that is taken at a different time, weeks, months, or even years later. Note that the localization problem does not mean a failure in localizing faces; rather, we are assuming that the face localization step succeeds, but that *small* errors of precision make the identification process fail. This problem has not attracted much attention to date.
- *Partially occluded faces*: One of the main drawbacks of the appearance-based paradigm (e.g., PCA), is its failure to robustly recognize partially occluded objects. Fig. 2 shows an example, where Fig. 2a is used for learning and Fig. 2b for testing.
- *Expression variant faces*: Yet another little-studied problem is the expression-invariant one. This problem can be formulated as follows: *"how can we robustly identify a person's face for whom the learning and testing*

• *The author is with the School of Electrical and Computer Engineering, 1285 EE Building, Purdue University, West Lafayette IN 47907. E-mail: aleix@ecn.purdue.edu.*

Manuscript received 20 Nov. 2000; revised 5 July 2001; accepted 24 Oct. 2001.
Recommended for acceptance by D. Kriegman.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 113174.

0162-8828/02/$17.00 © 2002 IEEE

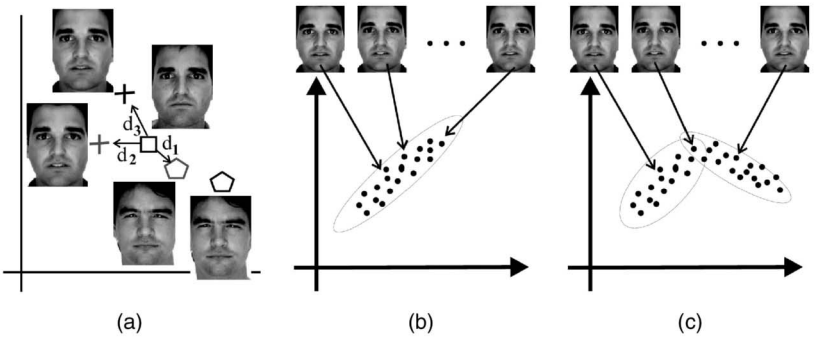


Fig. 1. (a) The localization problem: Different localization results lead to different representations onto the feature space, which can cause identification failures. (b) and (c) We learn the subspace where different localization lie.

face images differ in facial expression?" We say that a system is expression-dependent if the image of Fig. 3b is more difficult (or less difficult) to recognize than the image shown in Fig. 3c, given that Fig. 3a was used for learning. An expression-invariant method would correspond to one that equally identifies the identity of the subject independent of the facial expression displayed on the test image. In an attempt to overcome this problem, the PCA approach uses the second order statistics of the image set in the hope that these features will be invariant to different facial expressions. Unfortunately, it can be theoretically shown that, for any given invariant paradigm (the PCA approach in this case) there is always a set of (testing) images for which the learned measure (or function) will not be optimal [13], [64].

Every system that attempts to recognize faces in unconstrained (or close to unconstrained) domains has to deal with the above described difficulties.

One way to deal with these problems is by using large, representative training data sets. Large means that the number of images needed to achieve generalization must be proportional to the dimensionality of the feature space, which is a well-known fact within the pattern recognition (PR) [26], [24], [21], [62] and machine learning (ML) [65], [19], [42] communities. In PR, it is generally accepted that using at least 10 times as many training samples per class as the number of features is a good practice to follow [24]. The more complex the classifier, the larger this ratio should be [26]. Recall that, in the appearance-based domain, the feature space corresponds to a dense pixel representation, i.e., the number of dimensions equals the number of image pixels, which makes this problem more obvious (because the number of pixels of an image tends to be very large).

These samples also have to be representative, meaning that (nearly) every test image should have its corresponding training sample taken (roughly) under the same conditions. Otherwise, a test image can always be found which will not be correctly classified [64]. Unfortunately, reality falls far short of theory. The number of samples is usually not proportional to the dimensionality of the feature space (like 10 to one) and, worse, not even small (one-to-one) representative data sets are (generally) given. In many applications, only one or two samples per class are available to the system.

Martínez and Kak [35] have recently shown that the switch from nondiscriminant techniques (e.g., PCA) to discriminant approaches (e.g., LDA—Linear Discriminant Analysis) is not always warranted and may sometimes lead to faulty system design when small and nonrepresentative training data sets are used. In other words, the typical choice of the classification error criterion as the criterion function in dimensionality reduction might be inadequate when small, nonrepresentative sample sets are given.

In this article, we will define a representation of face images that is as invariant as possible to the above enumerated problems. Then, to make our system also robust to those changes that could not be characterized by our representation (feature space), we will propose methods that attempt to make the testing images as close as possible to the learning ones (or vice versa) before matching takes place. We summarize our approaches as follows:

- *Imprecise localization of faces*: We resolve this problem by learning the subspace that represents *small* localization errors within the eigenspace. We do this as follows: First, we need to define a way of finding this localization error (i.e., the physics to be modeled). (Notice that different localization algorithms will have different associated errors.) Once the localization error is known, we can synthetically generate all possible warped images and search for the subspace where these images lie. One way to do this is by modeling each subspace as a Gaussian

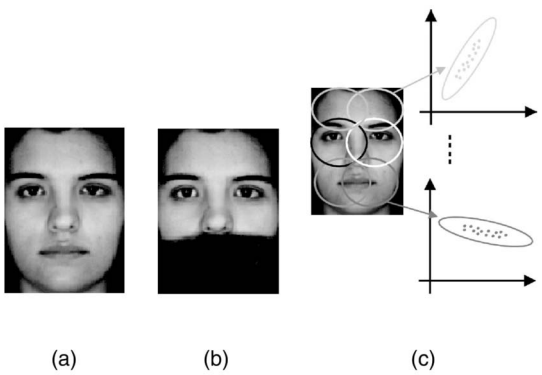


Fig. 2. (a) A possible learning image. (b) A test image. (c) The local approach.

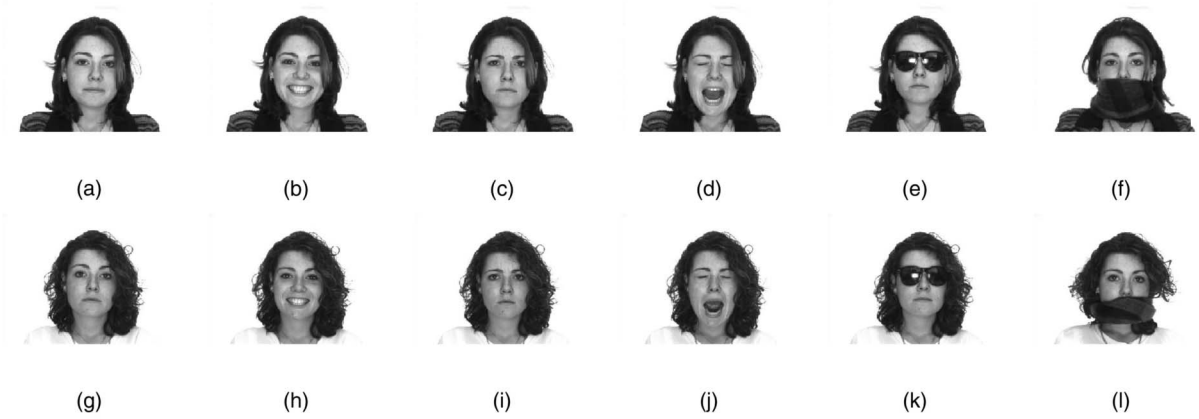


Fig. 3. Images of one subject in the AR face database. The images (a) through (f) were taken during one session and the images (g) though (l) at a different session (14 days later).

distribution (as it was introduced in [36]) or as a mixture of Gaussians.[1] This process is displayed in Figs. 1b and 1c for clarity.

- *Partial occlusions*. In order to cope with this problem, each face image is divided into k different local parts (as shown in Fig. 2c). Each of these k local parts is modeled by using the above-mentioned Gaussian distribution (or, equivalently, with a mixture of Gaussians), which accounts for the localization error problem. Given that the mean feature vector and the covariance matrix for every local subspace are now known, the probability of a given match can be directly associated with the sum of all k Mahalanobis distances. This approach differs from previous local PCA methods [7], [8], [43], [46] in that we use a probabilistic approach rather than a voting space.
- *Expression-invariant*. In general, different emotions are expressed as facial expressions with more emphasis on specific parts of the face than others [20], [55], [11], [22], [45]. Therefore, a face will be easier to identify at a given local area depending on the emotion currently expressed [37]. This knowledge can be added to the classifier as prior information by weighting each of the above k local parts with values that vary depending on the facial expression displayed on the testing image.

The rest of this paper is organized as follows: Section 2 describes the database and the localization and warping algorithms used. Section 3 reports on the localization problem (i.e., imprecise localization of faces), the physics to be modeled, and the approach to learn such physics. Section 4 introduces the occlusion problem and a probabilistic method that tackles it. Section 5 modifies that method to deal with changes of appearance due to differences in facial expression between the learning and testing images. We conclude in Section 6.

1. The idea of using synthetically generated images has been previously defined by several authors, e.g., [64], [5], [9], [66], [74]. However, the use of this technique to solve the *localization* problem, as well as the idea of learning the PDF that represents these synthetically generated images, is novel (to the author's knowledge).

2 The Database and the Localization Procedure

2.1 The AR-Face Database

The AR database of face images is publicly available at http://rvl1.ecn.purdue.edu/ARdatabase/ARdatabase.html [34]. This database consists of over 3,200 color images of the frontal view faces of 126 people. There are 26 different images per person, recorded in two different sessions separated by two weeks, each session consisting of 13 images. Only those images with frontal illumination are used in this communication (that is, a total of 12 images per person). For illustration, these images for one subject are shown in Fig. 3. All images were taken by the same camera under tightly controlled conditions of illumination. Viewpoint was also controlled (although some small 3D rotations might exist due to the nature of the images—see Section 2.3). Each image in the database consists of a 768×576 array of pixels and each pixel is represented by 24 bits of RGB color values. Since the number of frontal face images per person in the AR-face database is larger than in many of the other available face databases and since only one image is to be used for training purposes, we are now able to test the reported system under a large variety of conditions (e.g., occlusions and facial expressions).

2.2 Localization and Warping Algorithms

We describe the localization and warping methods in this section for two reasons: to allow reproduction of our results and to better describe both the localization error problem and the physics behind the method to be presented in Section 3. The reader should keep in mind that other alternatives for localizing faces (and facial features) exist, ranging from neural networks [54] to density estimation methods [59]. A recent survey is [72].

Color has now been largely used to classify objects [60] and to localize faces in an image [71]. Human skin color forms a dense manifold in color space, which makes it an easy feature to detect in images. We use the three-dimensional chromatic space (i.e., the normalized (R, G, B) space $(R/T, G/T, B/T)$, where $T = R + G + B$) to prelocalize faces in our images; Fig. 4b shows a result. The erosion and dilation morphological operators are used to remove small isolated segments,

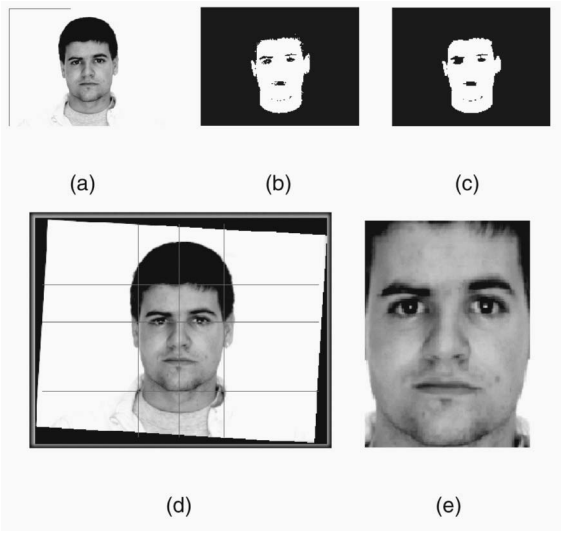


Fig. 4. Localization example.

Fig. 4c. The results reported in this communication use a 4×6 operator.

Then, the PCA technique as described in [43] is used to localize the eyes and the mouth of each face. To increase matching accuracy, one could use the approach proposed in [10] or some other feature-based methods such as [73], [31], [14]. Once these facial features have been localized, using the differences between the x and y coordinates of the two eyes, the original image is rotated until obtaining a frontal view face where both eyes have the same y value. Mathematically, $atan(\|y_1 - y_2\|/\|x_1 - x_2\|)$, where (x_1, y_1) and (x_2, y_2) are the right and left eye coordinates. Finally, a contour technique is used to search for the lateral boundaries of the face (i.e., left and right). The top and bottom of the face are assigned as a function of the above features. Fig. 4d depicts a face localization result. The face is then warped to a final "standard" face of 120 by 170 pixels, Fig. 4e. After warping, the eye centers, the medial line of the nose, etc., are expected to be at the same pixel coordinates in all images. The warping procedure is necessary to guarantee that every image pixel represents the same feature (this is known as the correspondence problem [5], [35]) and to enhance recognition (especially of duplicate images) as we describe below (in Section 2.3). The warping method described uses seven different facial features to achieve correspondence, i.e., the x and the y coordinates of the left eye, the x and the y coordinates of the right eye, the y coordinate of the mouth, and the x coordinate for both the left and the right edges (boundaries) of the face. This number, which we shall denote as f, could be increased to obtain better fits.

Later, we will also find it useful to specify the number of features that change their x coordinate value, which we denote as f_x, and the number of features that change their y coordinate value, which we denote as f_y. Note that $f = f_x + f_y$ and that, in our case, $f_x = 4$ and $f_y = 3$.

It is important to point out that some features cannot be detected in Figs. 3e, 3f, 3k, and 3l. To overcome this, the eye positions in Figs. 3e and 3k and the mouth position in Figs. 3f and 3l were obtained by averaging the values from all other face images (i.e., Figs. 3a, 3b, 3c, 3d, 3g, 3h, 3i, and 3j) of all subjects.

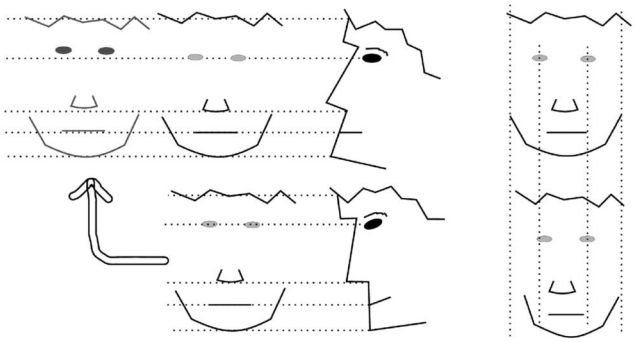


Fig. 5. The left part of this figure shows the shape changes that a face undergoes when tilted forward. The top image corresponds to a totally *frontal* view. The bottom image represents the projection of a slightly tilted face. The bottom face has been redrawn next to the top one, to facilitate comparison. The right part of this figure shows some of the visible deformations cause by rotating the face to the subject's right.

2.3 Why Is the Warping Step Necessary?

Most face recognition systems developed to date use 2D images to find the best match between a set of classes and an unclassified testing image of a face. It is important to keep in mind, however, that faces are three-dimensional objects and that, as such, their projection onto the 2D image plane may undergo complex deformations. The frontal face recording scenario (used in the AR-face database and others) is one of these cases and needs to be considered. Frontal face images are generally obtained by placing a camera in front of the subject who is asked to look at the camera while the picture is taken. The resulting 2D image projection obtained using this procedure is called a "frontal" face image. In general, several "frontal" faces are recorded, some to be used for training purposes and others for testing. However, the word "frontal" might be misinterpreted. Hence, faces are three-dimensional objects, their orientation with respect to a fixed camera is not guaranteed to be the same from image to image (especially for duplicates since people tend to orient their faces toward different locations on different days). For example, when the face is tilted forward, its appearance (shape and texture) in the 2D image plane changes. This effect is shown in the left part of Fig. 5. Tilting the face backward will result in other appearance changes. Also, rotating the face to the right or to the left will impose appearance changes; this is depicted in the right-hand side of Fig. 5.

These deformations make the identification task difficult because any combination of the above can be expected. Feature-based approaches might be inappropriate because the distance between both eyes, the eyes and the mouth, the mouth and the chin, etc., are not guaranteed to be stable within subjects. Texture-based approaches will also be affected because the facial features of a subject can easily change positions in the array of pixels used to identify the face. A pixel to pixel analysis is then inappropriate [5], [35]. Previous studies have also shown that better recognition rates are obtained when one uses a free-shape representation [67], [15].

To overcome this difficulty, one could use large and representative training sets. By doing this, one expects to collect enough learning data to sample the underlying distribution of the feature-space that represents all possible deformations described above. Unfortunately, there are so

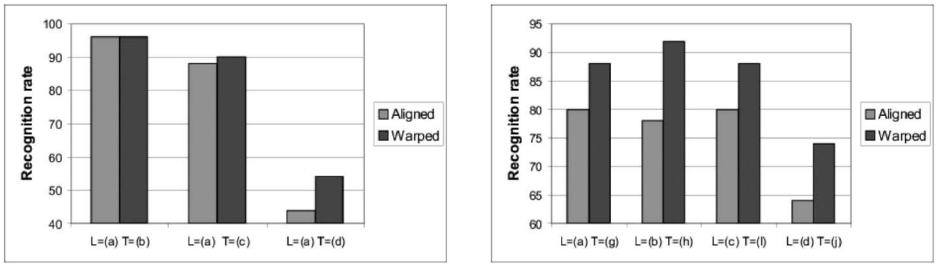


Fig. 6. Shown here are the recognition rates obtained by using either the aligned-cropped or the warped images.

many possible orientations of the face considered to be a "front viewed" that the collecting of this hypothetical training set might be impossible in most practical situations. When recognizing faces by means of a single learning image per class, there is no hope for the training data to underlie the true distribution of each subject.

To successfully tackle the above-defined problem, it is common to warp all faces to a "standard" shape (otherwise the PCA algorithm will also represent nondesirable shape features of the face [49]). The procedure described in the previous section is such a mechanism. Note, however, that the proposed warping algorithm of Section 2.2 does not *deform* the important facial features. The shape of the eyes, mouth, nose, etc., is *not* affected. This is important because these facial features may carry discriminant information. The warping mechanism only guarantees that the eyes, nose, and chin positions are at the same image coordinate for all individuals and that all images (independently of the facial expression displayed on the face) occupy a fixed array of pixels. This allows us to efficiently analyze the texture (appearance) of the face image and to use pixel to pixel differences. Caricaturing might improve the results a bit further [15].

Although, the above discussion is quite convincing with regard to the claim that the warping step is necessary to correctly use appearance-based methods and to successfully identify faces, we still bear the burden of establishing our claim with the help of real data. To do this, 50 subjects were randomly selected from the AR-face database. Images were warped to fit a 120×170 array of pixels by following the procedure described in Section 2.2. A second set of images, which we shall name the *aligned-cropped* set, was also obtained. For the latter set, each face was aligned only with regard to the horizontal line described by the eyes' centers; which requires rotation of the face image until the y values of both eyes' centers are equal in value (i.e., $atan(\|y_1 - y_2\|/\|x_1 - x_2\|)$, where (x_1, y_1) are the pixel coordinates of the center of the right eye and (x_2, y_2) is the same for the left eye). These faces were then cropped to fit an array of 120×170 pixels. After cropping, the horizontal that connects both eyes is at the same y pixel coordinate for all images.

A first (40-dimensional) PCA representation was generated by using the warped neutral expression face image of each of the 50 selected subjects. A second (40-dimensional) PCA space was generated by using the aligned-cropped neutral expression images of the same subjects. The neutral expression images (before warping or aligning-cropping) of one of the participants is shown in Fig. 3a. To test recognition accuracy, the happy, angry, and "screaming" face images (Figs. 3b, 3c, and 3d) of each participant were now used. The warped images were used to test the first learned PCA representation and the aligned-cropped face images to test the second. The results, as a function of successful recognition rate, are shown in Fig. 6a. In this graphical representation of the results, L specifies the image that was used for learning the PCA representation and T the one that was used for testing. For clarity, the results are shown separately for each group of testing images with a common facial expression. We see that the warping procedure helped to achieve higher recognition rates.

As mentioned earlier in this section, we should expect this difference to increase for duplicate images because, at different sessions, people tend to orient their faces differently. To show this effect, we used the neutral, happy, angry, and "screaming" images taken during the second session (Figs. 3g, 3h, 3i, and 3j) to test the PCA representations of their corresponding duplicate image taken at the first session (Figs. 3a, 3b, 3c, and 3d). The results are shown in Fig. 6b. Note that, as expected, the improvement in recognition is now much more notable.

3 Imprecise Localized Faces

3.1 Modeling the Localization Problem

It would be unrealistic to hope for a system that localizes all the facial features described above (e.g., eyes, mouth, etc.) with a high degree of precision. Due to different facial expressions, lighting conditions, etc., the localization stage is expected to be error prone. The problem this raises is that the feature representation (i.e., the feature vector) of the correct localized face differs from the feature representation of the actual computed localization, which can ultimately imply an incorrect classification result, as was visually depicted in Fig. 1a.

In order to tackle this problem, some authors have proposed filtering the image before attempting recognition, e.g., [30], [69], [70]. Filtered images are, in general, less sensitive to the localization problem, but do not fully tackle it. To impose further constraints, one can vary the weights of the vertices of an adequately defined graph that interconnects spatially adjacent local areas. Areas that have been localized with more uncertainty will have lower vertex values, indicating that these areas are not good candidates for matching. The *Dynamic Link Architecture* [30] and the *Elastic Bunch Graph Matching* approach [70] could be used to achieve this. The major disadvantage, however, would be that only the areas which have been reliably localized are used for recognition purposes. Ideally, one wants to use the information on all the areas of the face.

Fortunately, we can model the localization problem by means of a probability density function. Assume at first that the localization errors for both the x and y image axes (which we shall denote as vr_x and vr_y, respectively) are known. These two (error) values tell us that, statistically, the correctly localized face is within the set of images $\hat{\mathbf{x}}_i = \{\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,r}\}$, where i specifies the class (i.e., $i \in [1, n]$, with n being the total number of classes) and r the number of different possible localizations while varying the error value from $-vr_x$ to $+vr_x$ about the x image axis and from $-vr_y$ to $+vr_y$ about the y axis. We note that r increases exponentially as a function of f (the number of features localized in the face), i.e., $r = O(2^f)$. Precisely,

$$\begin{aligned} r &= \left((2vr_x + 1)\sum_{i=1}^{f_x}\binom{f_x}{i}\right)\left((2vr_y + 1)\sum_{i=1}^{f_y}\binom{f_y}{i}\right) \\ &= (2vr_x + 1)(2^{f_x} - 1)(2vr_y + 1)(2^{f_y} - 1), \end{aligned}$$

where f_x represents the number of features that change their x coordinate value and f_y the ones that change their y coordinate value (recall that $f = f_x + f_y$). Where f is large, this complexity is to be considered. In those cases, given that accounting for the error in only one feature at a time would not imply a significant change of appearance, a subset of all these r images is expected to suffice.

Once the data set $\hat{\mathbf{x}}_i$ has been generated, the subspace where each class i lies can be readily modeled by means of a Gaussian distribution,

$$\begin{aligned} \mu_i &= \frac{\sum_{j=1}^{r}\mathbf{x}_{i,j}}{r} \\ \Sigma_i &= (\hat{\mathbf{x}}_i - \mu_i)(\hat{\mathbf{x}}_i - \mu_i)^T, \end{aligned} \tag{1}$$

where μ_i and Σ_i are the mean feature vector and the covariance matrix of the Gaussian model i.

In some cases (especially, when the localization error is large), the subspace where the set of images $\hat{\mathbf{x}}_i$ lies cannot be assumed to be Gaussian. In such cases, a mixture of Gaussians is generally a better choice because it can approximate complex nonlinear subspaces [39]. Better fits come with considerable extra computational cost that must be taken into account. A mixture of Gaussians can be learned by means of the EM (Expectation-Maximization) algorithm [17], [40], which is an iterative method divided into two steps, the *E-step*

$$h_{i,j,g}^{[t+1]} = \frac{\left|\Sigma_{i,g}^{[t]}\right|^{-1/2} exp\left\{-\frac{\left(\mathbf{x}_{i,j}-\mu_{i,g}^{[t]}\right)^T\Sigma_{i,g}^{[t]^{-1}}\left(\mathbf{x}_{i,j}-\mu_{i,g}^{[t]}\right)}{2}\right\}}{\sum_{l=1}^{G}\left|\Sigma_{i,l}^{[t]}\right|^{-1/2} exp\left\{-\frac{\left(\mathbf{x}_{i,j}-\mu_{i,l}^{[t]}\right)^T\Sigma_{i,l}^{[t]^{-1}}\left(\mathbf{x}_{i,j}-\mu_{i,l}^{[t]}\right)}{2}\right\}} \tag{2}$$

and the *M-step*

$$\begin{aligned} \mu_{i,g}^{[t+1]} &= \frac{\sum_{j=1}^{r} h_{i,j,g}^{[t]}\mathbf{x}_{i,j}}{\sum_{j=1}^{r} h_{i,j,g}^{[t]}} \\ \Sigma_{i,g}^{[t+1]} &= \frac{\sum_{j=1}^{r} h_{i,j,g}^{[t]}\left(\mathbf{x}_{i,j} - \mu_{i,g}^{[t+1]}\right)\left(\mathbf{x}_{i,j} - \mu_{i,g}^{[t+1]}\right)^T}{\sum_{j=1}^{r} h_{i,j,g}^{[t]}}, \end{aligned} \tag{3}$$

where G is the total number of models used in the mixture (i.e., number of Gaussians) and $[t]$ means iteration t. In the above formulation, it is assumed that all models (Gaussians) are equiprobable (i.e., $\pi_{g_1} = \pi_{g_2}\ \forall g_1, g_2$, where π_g denotes the probability of the gth model). Alternatively, equations will differ slightly from above [40]. The initial values (i.e., $\mu_{i,g}^{[0]}$ and $\Sigma_{i,g}^{[0]}$) are taken at random.

The reader should keep in mind that several improvements can be applied to this standard EM algorithm for fitting mixtures of Gaussians as, for example, to avoid local maxima (e.g., [41], [53], [38]) or to determine a successful number of Gaussian models (e.g., [68], [52], [2]).

As an alternative to the above formulation, one could employ a nearest neighbor classifier based on the tangent distance of [56] or a deformable matching technique as in [25]. These methods carry the disadvantages of being invariant only to small deformations of the sample data and of being computationally very expensive. The computational cost might be overcome by using the *tangent-prop* algorithm [57]. The tangent-prop algorithm incorporates the tangent direction of each training sample within the backpropagation algorithm to learn the function that describes the object (e.g., face) to be learned. This allows better approximations of the function to be learned than when one only uses the backpropagation algorithm. The tangent values can be readily computed by synthetically deforming the samples to be used in the learning process slightly (which allows the computation of the discrete derivatives of the image plane). This algorithm might not be able to learn the underlying manifold as precisely as the algorithm described above, but would be computationally less demanding. It has been shown that, as the number of samples increase, the tangent-prop algorithm reduces its error-on-test-data quicker than the backpropagation algorithm for the recognition of handwritten digits [57].

3.2 The Eigenspace Representation

Since the feature space corresponds to a dense (pixel) feature representation, its dimensionality is expected to be too large to allow the computation of the subspaces described above (Gaussians or mixture of Gaussians). For example, if face images are 100 by 100 pixels, the dimensionality of the feature spaces will be 10,000 dimensions. From our previous discussion on the minimal number of samples needed to successfully estimate the parameters of our model (like 10 to one), it follows that 100,000 or more samples are needed; obviously a prohibitive number.

In a way, it could be said that the localization problem solves (in some cases) the curse of dimensionality. For example, if $f = 10$ (say, $f_x = f_y = 5$) and $vr_x = vr_y = 5$, then $r = 123{,}783$. However, we recall that all these samples are highly correlated and thus, unfortunately, cannot be considered as independent training images. In fact, only one vector (image) per sample is generally (statistically speaking) significant.

Even if, for some cases, the use of all these samples seems to be best suited (because, for example, the subspaces that they describe seem only to involve a group of specific dimensions of the feature space), the computational cost might not allow us to use all of them. We note that, even for the single Gaussian case, d^2 sized matrices have to be computed (where d is the dimensionality of the feature space). If d is large, say 10,000 dimensions, as much as 95.3674 Mbytes are required to store each Gaussian model (assuming that only one byte is used to represent each value of the matrix).

The above discussion leads to the conclusion that only one warped image per sample should be considered when determining the (top) dimensionality of our feature space. In other words, we need to reduce the dimensionality of the original feature space to one where the number of samples per dimension is appropriate. This is done by means of the PCA approach. Once the eigenspace has been obtained, all r samples are projected and the Gaussian models (or, equivalently, the mixture of Gaussians) are computed.

Mathematically, we generate the eigenspace by taking the e eigenvectors associated to the largest eigenvalues of $\mathbf{Q} = (\mathbf{X} - \mu)(\mathbf{X} - \mu)^T$, where $\mathbf{X} = \{\mathbf{x}_{1,1}, \ldots, \mathbf{x}_{n,1}\}$, $\mathbf{x}_{i,1}$ corresponds to the first (original) image sample of class i in its vector form, and μ represents the mean value of $\mathbf{X}$. We are assuming that all faces have been previously localized and warped to the standard shape, as described in Section 2.2. Once these eigenspaces have been created, we can project the data sets $\hat{\mathbf{x}}_i \; \forall i$ onto the eigenspace and search for the Gaussian model or mixture of Gaussians that represents them.

In order to identify a new previously unseen image $\mathbf{t}_z$, we search for the closest model $\mathcal{G}_i$ (where $\mathcal{G}_i = \{\Sigma_i, \mu_i\}$) by means of the Mahalanobis distance in the reduced eigenspace

$$Mh^2(\hat{\mathbf{t}}_z, \mathcal{G}_i) = (\hat{\mathbf{t}}_z - \mu_i)\Sigma_i^{-1}(\hat{\mathbf{t}}_z - \mu_i)^T, \tag{4}$$

where $\hat{\mathbf{t}}_z$ is the projection of the warped (testing) image onto the eigenspace. For the mixture of Gaussians case, we have $\mathcal{G}_i = \{\Sigma_{i,g}, \mu_{i,g}\}_{g=1}^{G}$. When $G > 1$, the mean distance from the test feature vector to all G Gaussian models is taken as the similarity measure.

3.3 What Is the Value of r?

Recall that we still need to know the error that our localization algorithm is performing, that is to say, the physics to be modeled, i.e., the value of r. Obviously, this is a difficult task that, unfortunately, does not have an analytical solution for most face localization algorithms.

If the correct localization values are known (i.e., the ground-truth is known) for a set of s samples, $\mathbf{x}(s) = \{\mathbf{x}_1, \ldots, \mathbf{x}_s\}$, an estimation $E(r; \mathbf{x}(s))$ can be computed which depends on the number of samples s. It is easy to show that $E(r; \mathbf{x}(s))$ approximates the true value of r as $s \to \infty$. Obviously, we do not have that much data and, thus, only estimations can be made. The reader should keep in mind, however, that as s increases, $E(r; x(s))$ progressively approaches the true value of r. In general, the correct location of every feature has to be manually determined, which is also a cost to be considered.

For the experimental data reported here, we manually localized each of the above enumerated facial features for the images of 50 individuals. In doing that, we took care to be as precise as possible. Then, we computed the error of the localization algorithm described in Section 2.2 by comparing its results to our manually localized features. The localization was found to have a variance error of ± 3 pixels about the x image axis and of ± 4 pixels about the y image axis; $vr_x = 3$, $vr_y = 4$. For our particular case with $f_x = 4$ and $f_y = 3$, we have $r = 7 \cdot (2^4 - 1) \cdot 9 \cdot (2^3 - 1) = 6,615$.

3.4 Experimental Results

Experiment 1. In this test, we want to answer the following question: Assuming that the localization error is precisely known, how well does the above method solve the localization problem? For this purpose, we (manually) ground-truth the neutral expression, happy, angry, and "screaming" face images (Figs. 3a, 3b, 3c, and 3d) for 50 randomly selected individuals of the AR-face database. In doing that we took care to be as close to pixel precision as possible. Then, we added artificial noise to these data. In such a way, the values of vr_x and vr_y are known with precision. Since the localization error is now known, the recognition rate is not expected to drop as a function of vr_x and vr_y.

To test this, we use the neutral expression image (Fig. 3a) of all 50 subjects for learning and all others (Figs. 3b, 3c, and 3d) for testing. Results for vr_x and vr_y equal to 2, 4, 6, 8, and 10 are shown in Fig. 7a for the single (GM) and multiple Gaussian (MGM) case with $G = 3$. For simplicity, both error values were made to be equal. Results are compared to the classical eigenspace approach that does not account for the localization error. For the eigenspace approach, the Euclidean distance is used to find the closest match (class). In the results reported here, $e = 20$.

For vr_x and vr_y values up to 4, neither of the new methods is affected by the localization error. The multiple Gaussian case obtained slightly better results (as expected) and was not affected by the localization error problem until vr_x and vr_y were set to 8. Obviously, the larger the localization error, the more probable that two Gaussian models (or mixture of Gaussians) overlap. For our particular case of 20 dimensions and 50 people, this happened at $vr_x = vr_y = 6$ for the single Gaussian case and at $vr_x = vr_y = 8$ for the multiple Gaussian case with $G = 3$.

Moreover, note that results of the new proposed methods are superior to the classical PCA approach. This can be explained by the following: In an attempt to overcome the localization error problem, the Gaussian or mixture of Gaussians distributions give more importance (weights) to those combination of features that best describe small deformations of the face (small changes of the texture of the face image). It is reasonable to expect to also find these combinations of features more adequate for the recognition of testing face images that express different facial expressions, because the testing images also represent texture changes of the face. It is obvious that these features will by no means be optimal, but they can explain the small increase of recognition rate obtained in our experiment.

It is also interesting to show how the new proposed method behaves as a function of *rank* and *cumulative match score* in comparison to the classical PCA approach (which serves as baseline). *Rank* means that the correct solution is within the R nearest neighbors and *cumulative match score* refers to the percentage of successfully recognized images [51]. Fig. 7b shows results for $vr_x = vr_y = 4$ and $vr_x = vr_y = 10$ for case of G = 1. We note that, even for localization errors as large as 10 the recognition results do not diverge much from the classical PCA approach (i.e., the $v_x = v_y = 0$ case).

Experiment 2. We still need to see how well our method performs when the localization error is not known with precision. In this experiment, the localization method described in Section 2.2 is used. An approximation for the localization error was obtained in Section 3.3, $vr_x = 3$ and $vr_y = 4$. Again, the neutral expression image (Fig. 3a) is used for learning, while the happy, angry, and "screaming" faces (Figs. 3b, 3c, and 3d) are used for recognition. Fig. 7c shows the recognition results for the group of 50 people used to obtain the values of vr_x and vr_y of Section 3.3. Fig. 7d shows the results obtained with another (totally different) group of

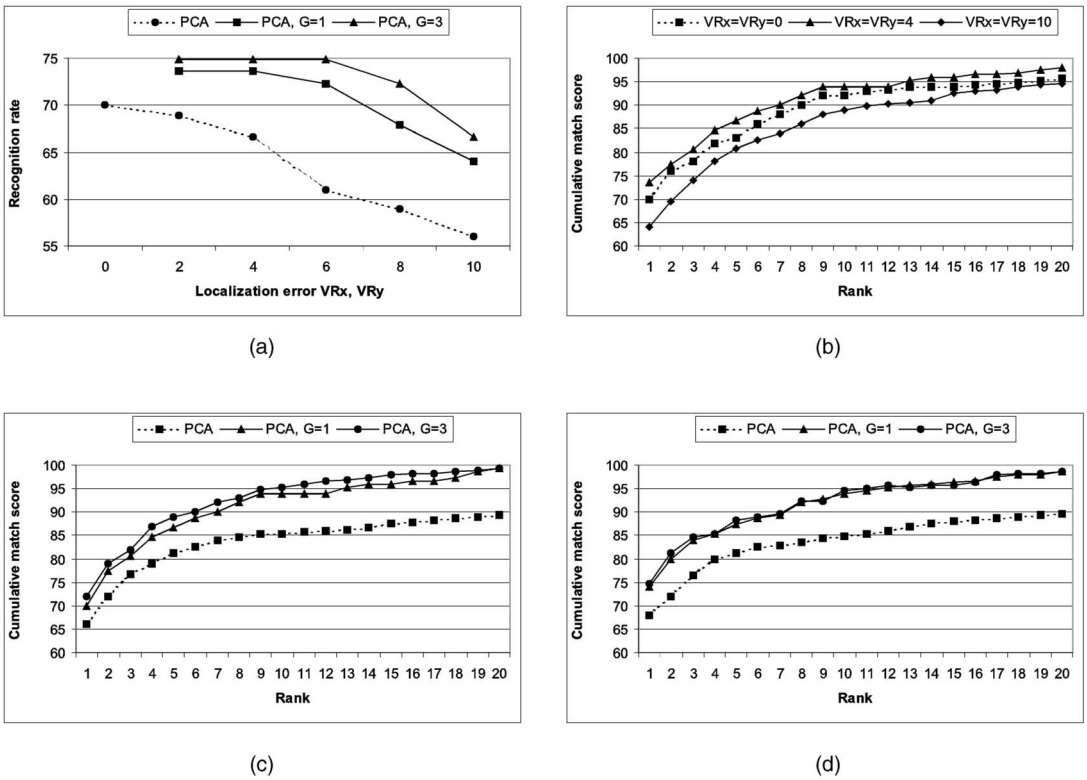


Fig. 7. This figure summarizes the results of *Experiment 1*, which are displayed in (a) and (b), and the results of *Experiment 2*, which are shown in (c) and (d).

50 people. Since the values of vr_x and vr_y were obtained from the first group of 50 people, the results of Fig. 7c are expected to be better than those of Fig. 7d. Moreover, it seems to be the case (as for the data shown here) that the more uncertainty we have in the estimation of the localization error, the more independent our results are from G.

It is also expected that the results of the proposed method will approximate the baseline PCA as the uncertainty of vr_x and vr_y increases. In order to show this, we repeated the first part of this experiment (*Experiment 2*) three more times with $vr_x^{[t+1]} = vr_x^{[t]} - 1$ and $vr_y^{[t+1]} = vr_y^{[t]} - 1$, where $vr_x^{[t]}$ and $vr_y^{[t]}$ represent the localization error values estimated at time t. We initially set $vr_x^{[0]} = 3$ and $vr_y^{[0]} = 4$ (values obtained in Section 3.3). Results are summarized in Table 1, with $G = 1$ and $e = 20$. As expected, the smaller the error values are made, the closer the results of our method are to results of the baseline PCA algorithm.

4 Partially Occluded Faces

One way to deal with partially occluded objects (such as faces) is by using local approaches [7], [8], [23], [43], [46], [36]. In general, these techniques divide the face into different parts and then use a voting space to find the best match. However, a voting technique can easily misclassify a test image because it does not take into account how good a local match is. Also, when large, representative data sets are available for learning, other local techniques—such as the *Local Feature Analysis* described in [48]—could be used. Here, we present a probabilistic approach that is able to compensate for partially occluded faces from a single training sample per class.

4.1 Learning Stage

In order to overcome the occlusion problem, we first divide each face image into k local parts. Then, we compute the method described above for each of the local parts. This implies the construction of k eigenspaces with their associated Gaussian (or mixture of Gaussians) distributions. More formally,

$$\mathbf{X}_k = \{\mathbf{x}_{1,1,k}, \ldots, \mathbf{x}_{n,1,k}\}, \tag{5}$$

where $\mathbf{x}_{i,j,k}$ is the kth local area of the (warped) sample image of class i in its vector form, $j = 1$ means the original training image (that is to say, only one image per subject is used, namely, the given sample because we are not accounting for the localization problem yet), and n is the total number of classes (i.e., people). We assume that all learning face images have been previously localized and warped to a standard shape as described earlier. Unless otherwise stated, we also assume $k = \{1, 2, 3, 4, 5, 6\}$, i.e., six local areas. To obtain each of the samples $x_{i,1,k}$, ellipse-shaped segments as defined by

TABLE 1
As the Values of vr_x and vr_y Are Made Smaller, the Recognition Rates Obtained Progressively Approach the Baseline Value of the PCA Algorithm

	t = 0	t = 1	t = 2	t = 3
Recognition rate	70%	68.3%	66.6%	66%

$x^2/d_x^2 + y^2/d_y^2 = 1$ are used (Fig. 2c), with each segment described in its vector form. After finding the mean vector μ_k of each $\mathbf{X}_k$, the covariance matrices $\mathbf{Q}_k = (\mathbf{X}_k - \mu_k)(\mathbf{X}_k - \mu_k)^T, \forall k$ are computed. Eigenspaces are obtained by taking the e eigenvectors associated with the largest eigenvalues of $\mathbf{Q}_k$. In the following, we shall refer to these eigenspaces as $\mathcal{E}ig_k$ and to their projecting matrices as $\mathbf{E}_k$ (which are constructed by means of the e eigenvectors associated to the largest eigenvalues of $\mathbf{Q}_k$).

Once all those eigenspaces have been computed, we are ready to search for the subspaces that account for the localization error. As described in the previous section, we now need to project all learning samples (accounting for the localization error) onto the above-computed eigen-representation. Mathematically speaking, we define $\hat{\mathbf{X}}_k = \{\hat{\mathbf{x}}_{1,k}, \ldots, \hat{\mathbf{x}}_{n,k}\}$, where $\hat{\mathbf{x}}_{i,k} = \{\mathbf{x}_{i,1,k}, \ldots, \mathbf{x}_{i,r,k}\}$ and represents all possible images accounting for all possible errors of localization. For obvious reasons, each $\hat{\mathbf{X}}_k$ is only projected onto its corresponding eigenspace by means of $\mathbf{E}_k$. Each $\hat{\mathbf{x}}_{i,k}$ set is expected to be within a small subspace of its corresponding $\mathcal{E}ig_k$, which can be modeled by means of a Gaussian distribution $\mathcal{G}_{i,k}$ with an associated mean $\mu_{i,k}$ and covariance matrix $\Sigma_{i,k}$, where $\mathcal{G}_{i,k}$ is the Gaussian model associated with the training sample $\hat{\mathbf{x}}_{i,k}$. As above, we shall denote the general case of single and multiple Gaussians as $\mathcal{G}_{i,k} = \{\Sigma_{i,k,g}, \mu_{i,k,g}\}_{g=1}^{G}$ (with $G = 1$ for the single Gaussian case). Notice that the set $\hat{\mathbf{x}}_{i,k}$ (which can be very large) does not need to be stored in memory; only the Gaussian (or mixture of Gaussians) model is needed for consecutive computations.

4.2 Identification Stage

When a test image $\mathbf{t}_z$ is to be recognized, we work as follows: We first localize the face and warp it to its final 120×170 pixel array by means of our earlier described method. This we denote as $\bar{\mathbf{t}}_z$.

We project each of the six local areas onto the above computed eigenspaces. More formally, $\hat{\mathbf{t}}_{z,k} = \mathbf{E}_k \cdot \bar{\mathbf{t}}_{z,k}, \forall k$, where $\bar{\mathbf{t}}_{z,k}$ represents the kth local part of $\bar{\mathbf{t}}_z$, and $\hat{\mathbf{t}}_{z,k}$ its projection onto $\mathcal{E}ig_k$. Since the mean feature vector and the covariance matrix of each local subspace are already known, the probability of a given local match can be directly associated with a suitably defined distance

$$LocRes_{i,k} = (\hat{\mathbf{t}}_{z,k} - \mu_{i,k})\Sigma_{i,k}(\hat{\mathbf{t}}_{z,k} - \mu_{i,k}). \quad (6)$$

For the mixture of Gaussians case, we compute the mean probability of the mixture. Finally, we add all local probabilities,

$$Res_i = \sum_{k=1}^{6} LocRes_{i,k}, \quad (7)$$

and search for the maxima,

$$ResClass = argmax_i\ Res_i, \quad (8)$$

where $ResClass \in [1, n]$.

If a video sequence is supplied, we keep adding distances (probabilities) for each of the images and only compute (8) at the end of the sequence or when a threshold has been reached.

When a local area is mostly or totally occluded, it is theoretically incorrect to project this area onto the corresponding eigenspace representation and search for the local match there. If an image does not correspond to a face, its projection onto the eigenface-space will be meaningless from an identification point of view (actually, we will be computing how close the original image is to our face-space). The problem now is to be able to automatically determine when an area is "enough" occluded to not be considered part of the face. This problem can be reformulated as a face detection problem, for which many algorithms exist [72]. In this communication, we do this by means of the eigenface representation, as described in [59]. For a given local area, we compute the distance to the eigenspace and then, within this eigenspace, we compute the distance to the mixture of Gaussians that represents this local area of the face.

4.3 Experimental Results

Experiment 3. The first question we want to address is: *How much occlusion can the proposed method handle?* To answer this question, the following procedure was followed: The neutral expression images (Fig. 3a) of 50 randomly selected participants were used for training the above defined system. For testing, synthetic occlusions were added to the (above-used) neutral images and to the smiling, angry, and screaming face images (Figs. 3a, 3b, 3c, and 3d). The occlusion was simulated by setting all the pixels of a square of size $p \times p$ pixels to zero. We tested values of p from a low of 5 to a maximum of 50. For each of these values of p, we randomly localize the square in the image 100 times (for each of the four testing images, i.e., neutral, happy, angry, and "screaming"). Fig. 8a shows the mean and standard deviation of these results for each of the facial expression groups. Fig. 8b shows the results one would obtained with the classical PCA approach. To increase robustness, one can increment the number of local areas (i.e., k). However, as k increases, the local areas become smaller and, thus, the method is more sensitive to the localization problem. At the extreme case, k will equal to the number of pixels in the image, i.e., our method will reduce to a correlation approach. In this case, only those pixels that have not been occluded will be used to compute the Euclidean (norm 2) distance between the testing and training images. The results obtained by means of this approach are shown in Fig. 8c. For comparison, we show the results obtained using the classical correlation algorithm in Fig. 8d. In general, the most convenient value of k will be dictated by the data or application.

Experiment 4. The experiment reported above was useful to describe how good our algorithm is when identifying partially occluded faces for which the occlusion size and location are not known a priori. Moreover, it is also interesting to know the minimum number of local areas needed to successfully identify a partially occluded face. To study this, the neutral expression images (Fig. 3a) of 50 individuals were used for learning, while the smiling, angry, and "screaming" images were used for testing (Figs. 3b, 3c, and 3d). Four different groups of occlusions were considered, denoted as occ_h with $h = \{1, 2, 3, 4\}$. For $h = 1$, only one local area was occluded, for $h = 2$ two local areas were occluded, etc. Occlusions were computed by discarding h of the local areas from the calculation. Given that the face is divided into k local areas, there are many possible ways to occlude h local parts. For example, for $h = 1$ there are k possibilities (recall that in our particular case $k = 6$), for $h = 2$, we have $\sum_{q=1}^{k}(k - q)$ possibilities, etc. For each value of h, all possibilities were tested and the

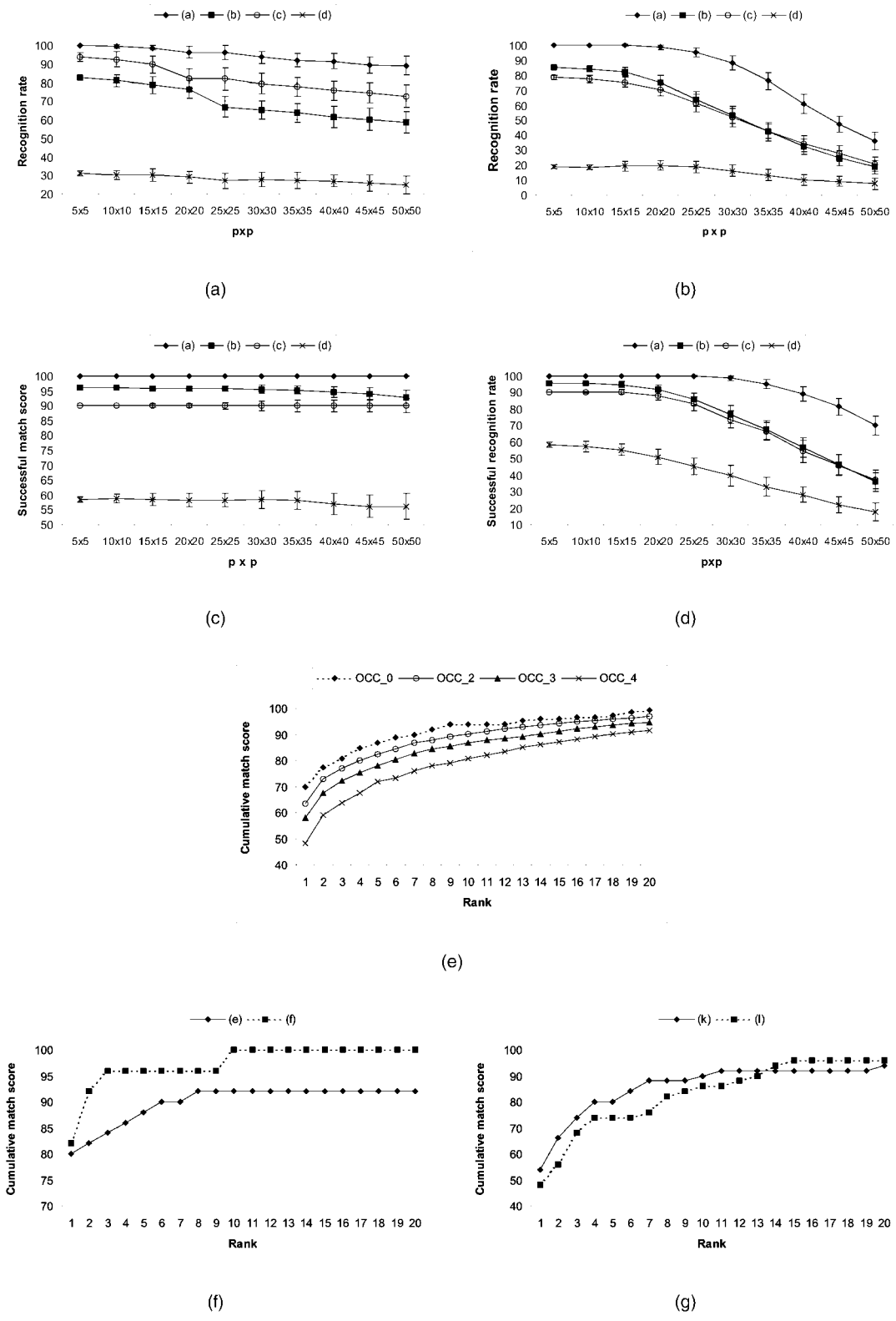


Fig. 8. Experimental results for (a), (b), (c), and (d) *Experiment 3*, (e) *Experiment 4*, and (f) and (g) *Experiment 5*.

mean result was computed. Fig. 8e shows the results. occ_0 serves a baseline value (which represents the nonocclusion case). We have not included the results of occ_1 in this figure because its results were too similar to those of occ_0 to be visually distinguishable. We conclude from this that, for our experimental data at least, the suppression of one local part did not affect the recognition rate of the system. The results of occ_2 reflect that our method is quite robust, even for those cases where a third of the face is occluded. For this test, $e = 20$ and $G = 1$.

Experiment 5. The next question we are interested in is: *How well does our method handle real occlusions?* For this purpose, we study two classical wearing occlusions, the sunglasses and the scarf occlusions (Figs. 3e and 3f). The neutral expression images (Fig. 3a) were used for learning, while the occluded images (Figs. 3e and 3f) were used for testing. We used the automatic localization algorithm described in Section 2.2, $vr_x = 3$ and $vr_y = 4$ (as described in Section 3.3), $e = 20$ and $G = 1$. Fig. 8f displays the results as a function of *rank* and *cumulative match score*. As a final test, we repeated this experiment (*Experiment 5*) for the duplicate images, using the neutral expression images of the first session (Fig. 3a) for learning and the occluded images of the duplicates for testing (Figs. 3k and 3l). Fig. 8g shows the results. It seems obvious from Fig. 8f that the occlusion of the eyes area affects the identification process much more than the mouth occlusion. This result is to be expected because it is believed that the eyes (or eyes' area) carry the most discriminant information of an individual's face. Surprisingly, though, this was not the case for the recognition of duplicates, Fig. 8g. The occlusion of the eyes area led to better recognition results than the occlusion of the mouth area. It seems to be the case that little is still known about how the recognition of duplicate works (or can be made to work). Further studies along these lines are needed.

5 Expression-Invariant Recognition

Results reported in the literature for frontal face images with as few as 50 individuals are normally about 90 percent correct or above. In this article, we report results of the PCA algorithm below 70 correct percent (for rank = 1), as shown in Figs. 7a and 7b. Our method obtained identification results around 75-85 percent correct (Fig. 7a and Fig. 8). This discrepancy is due to the fact that the databases used in the literature are relatively simple in comparison to the AR-face database. Note, for example, the extreme change in texture between the image displayed in Fig. 3a and the one shown in Fig. 3d. Most of the systems described in the literature are tested only with images that have similar facial expressions to those shown on the training image(s). In fact, as discussed in the introduction, this is the correct way of doing this from a pattern recognition point of view. Unfortunately, this does not reflect a general, realistic application, where large changes of expression can be expected. Also, even when the facial expression does not change, different orientations of the face can make different facial expressions "appear" in the 2D images displayed on the CCD of the camera. For example, it has been shown [33] that there is a relationship between the vertical angle from which a 3D neutral expression face image is viewed (from a fixed camera) and the judgment of facial expression that is viewed in the resulting 2D display. Faces that are tilted forward are judged as happier, faces tilting backward as sadder. In other words, the point of view from which one looks at a 3D face also affects the texture obtained in the 2D array of pixels.

The question of how much each of the different facial expressions included in the AR-face database affects the recognition rate is addressed in Fig. 9. To obtain the results displayed in this figure, we used the neutral expression images (Fig. 3a) from a total of 50 individuals for training, and the happy, angry, and "screaming" faces (Figs. 3b, 3c, and 3d) for testing. As expected, the more a testing face image differs from the training sample, the worse the recognition rate is. In

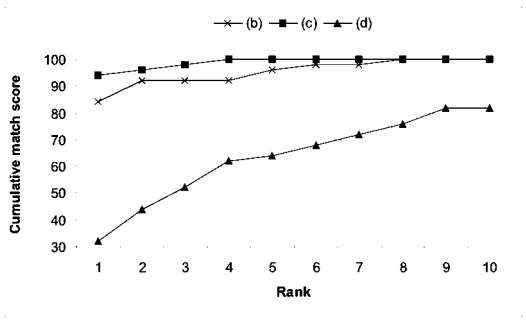


Fig. 9. The recognition rate obtained when using appearance-based approaches strongly depends on the difference between the facial expression displayed on the learning and testing images.

Fig. 10, we display the differences (i.e., pixel to pixel subtraction) between the neutral expression face image of a subject and her angry, happy, and "screaming" faces. The last image (Fig. 10i) is the difference between the neutral expression image taken at the first session and its duplicate (i.e., her neutral expression image at the second session). Note that the "screaming" face (Fig. 10d) differs much more from the neutral expression than the duplicate image (Fig. 10e).

5.1 Weighting Local Areas

In this section, we extend the previously introduced probabilistic approach so that it is less sensitive to the facial expression displayed on the testing image. The construction of a system that completely overcomes this problem may not even be realistic as it is known to be a very difficult task for humans too. Fortunately, some knowledge about the way one expresses a certain emotion (through one's facial muscles) can be used as prior information in an attempt to make the recognition system less sensitive to the differences between the texture patterns displayed on the learning and testing face images.

One way to cope with the above difficulty would consist of morphing all testing images to equal (in shape) the learning one [64]. In order to achieve this, we can compute the optical flow between the learning and testing images and then use this motion information to morph the test face [5], [6] to equal in shape the one used for training. Unfortunately, this cannot always be achieved, e.g., the eye area of a "screaming" image (Fig. 3d) cannot be morphed to a neutral eyes expression because, in most of the cases, the texture of the inside of the eyes is not available and cannot be inferred from the training data [5]. To solve this, we might try to build a system that synthetically generates all possible facial expressions (in a similar way as we tackled the localization problem). Unfortunately, here, the way one person expresses an emotion is (generally) somehow different from the way another subject expresses the same emotion (see Fig. 11 for an example) and, therefore, an optical flow technique would generally be inappropriate for the job.

Also, one could use the tangent-prop algorithm [57] described above to now learn the local directions within the feature space that determine the expression changes of each person's face. This approach, however, would have the same problem described in the previous paragraph because samples tend to be grouped by facial expression rather than by identity in the correlation space [1].

Fig. 10. Differences between the neutral expression image of a subject and her happy, angry, "screaming," and duplicate face images. Lower row is as: $f = b - a$, $g = c - a$, $h = d - a$, and $i = e - a$.

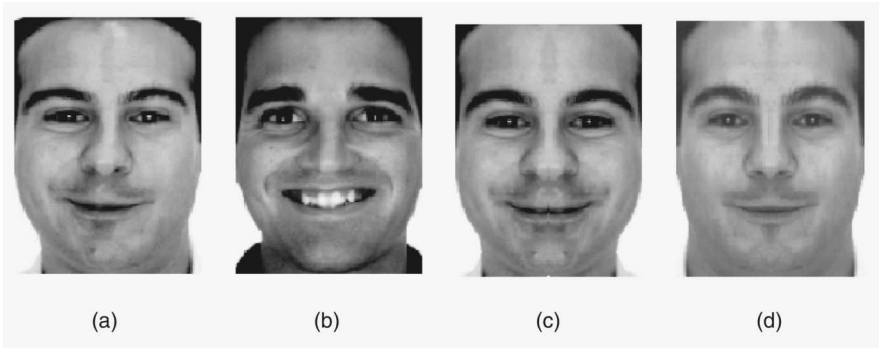
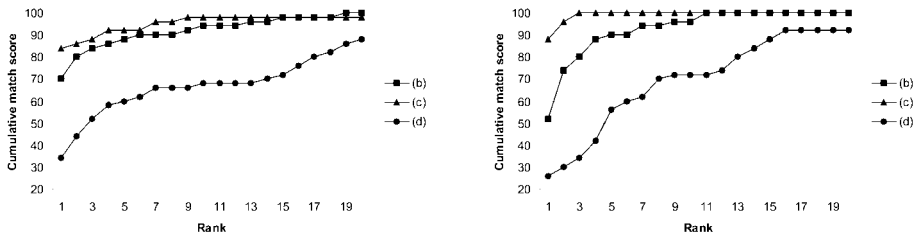


Fig. 11. (a) and (b) Although smiles (happiness) are basically the same across individuals, differences exist which make current "optical flow" morphing methods inappropriate for simulating emotions on a person's face. (c) Composite of the left part of the image shown in (a) and its mirror image. (d) composite of the right part.

Another way to circumvent the expression difficulty is to analyze which areas of the face tend to change more for a given emotion. It was proposed by Sackheim et al. [55] that emotions (i.e., facial expressions) are most expressed on the left side of the face. If true, our recognition system might best analyze the right side of the face (because this is the one that diverges less from the neutral face image). Critics of this type of experiment have noted that such studies fail to distinguish between photographs of posed emotions and those of genuine emotions [22]. This asymmetry seems to be less evident for spontaneous, genuine emotions. New experiments provide a better explanation. Since anger and sadness involve the right hemisphere more than the left, whereas happiness involves the left, it is logical to assume that anger will be more expressed on the left side and happiness on the right. It has been shown that composite photos constructed of the left side of the face are judged as happier, and composites of the right side of the face are judged as sadder [11]. A clear case of asymmetry in facial expression is seen in Fig. 11a. For this image, we also show the composite photos of the left part of the face, Fig. 11c, and of the right part of the face, Fig. 11d. Do you agree with most people in that the image shown in Fig. 11c looks happier than the one in Fig. 11d?

In order to determine whether the above statement can be useful, a new test was pursued. This consisted of identifying faces based only on the left or the right local areas among the six local parts described earlier. This results in two cumulative-match-score/rank curves per facial expression group, Fig. 12. It is easy to see that for those face images expressing: 1) happiness, the left side of the face gives better results than the right, 2) anger, the right side achieves better recognition results than the left, and 3) complex expressions (e.g., "screaming") do not have a clear preferred side.[2] A

2. For this latter case (that is to say, the "screaming" face), one could argue that the results obtained on the left side are better than the ones obtained on the right. This would be explained on the basis that the "screaming" face can be viewed as an extreme case of the happy (smile) display. However, the recognition rates we obtained in our experiment are too low to favor such a conclusion. Also, the results obtained on the left side of the face are not statistically significant when compared to those obtained on the right side.



(a) (b)

Fig. 12. Recognition rates obtained by only using (a) the three left local areas and (b) the three right local areas.

simple weighting approach would consist of giving more importance to the results obtained from the left local areas when the face to be identified expresses happiness and the opposite where the expression displayed on the testing image is sadness or anger. Equal values could be given to those emotions which equally affect both sides of the face or when a preferred side cannot be chosen.

It is also known that different facial expressions influence different parts of the face more than others [20], [16]. For example, whereas anger is (generally speaking) more emphasized in the eyes and the mouth areas, happiness also involves the cheeks (which rise) and nose. This information can now be included in the classifier as prior information. One way to do this would be to analyze which of those areas of the face are (in general) most expected to change for a given facial expression and then weight each local area accordingly. Unfortunately, there is no consensus on which of the areas are most affected by a given facial expression and why [29]. Furthermore, small differences in the way subjects express emotions exist due to cultural differences. Although this last statement is mainly true for those cases where large cultural changes exist (e.g., western versus oriental cultures), we cannot skip this fact when designing our classifier.

In order to resolve these problems, we built a learning mechanism that searches for those areas that are less affected by a given emotion within a group of people. Once these weights have been obtained, we can use them within the same group of individuals or for a different group from a culture that does not diverge much from the first group of people. Where large cultural differences exist, these weights should be relearned. It is important to note though that, in practical applications, those large cultural differences do not (normally) exist.

Mathematically, we can easily define the weighting approach as

$$Res_i = \sum_{k=1}^{6} w_{j,k} LocRes_{i,k}, \tag{9}$$

where $w_{j,k}$ are the weights associated to each of the six local areas $k \in [1,6]$ given a facial expression $j \in [1,m]$. These weights are to be learned from some training data. Notice that this implies the set of learning data to be labeled (not only with identity, but also with facial expression). We also note that the facial expression of the testing face needs to be known. Fortunately, several techniques exist to accomplish this task [18], [32], [61], [47].

The weights can be determined as follows:

$$w_{j,k} = q_{j,k} / \sum_{k=1}^{6} q_{j,k}, \tag{10}$$

where $q_{j,k}$ is the number of images with expression j successfully identified using only the kth local area. It is important to keep in mind that the data used for training these weights (for example, the one used earlier in the experimental results of Section 3) can never be the same used for testing purposes. Also, note that learning is now divided into two steps (as depicted in Fig. 13). In the first stage, the weights are set with the help of a data set adequately labeled by identity and facial expression. In the second step, the weighted eigenspace is built by means of the samples that correspond to the new individuals (for which the system is going to be later tested) and the previously obtained weights. The closest class is found by means of (8), as before.

Similarly as defined above, one could use the *Dynamic Link Architecture* [30] to determine which of the k local areas of the face have undergone the strongest deformation and weight each area accordingly. This approach would be adequate for images that have local changes (e.g., happy faces), but would not be adequate for those images with expressions that affect the whole face (e.g., "screaming" faces). The advantage of the method described above is that it learns which of the local areas carry the most discriminant information independently of how much (or how little) the face has been locally deformed. Also, an alternative way to compute the above weights might be by means of the *gating network* approach [27].

5.2 Experimental Results

The two last experiments were designed to evaluate the above system for those cases where the group of people does not change (i.e., although the images used for learning and testing are different, they all belong to the same individuals) and where the group of people does change (but where large cultural differences between groups do not exist).

Experiment 6. For learning, the neutral expressions image (Fig. 3a) is used to compute the eigenspace and to learn the subspaces that account for the localization error (same as in Sections 3.4 and 4.3). The other images (Figs. 3b, 3c, and 3d) are used to learn the weights associated to each local area, $w_{j,k}$, $j = \{1,2,3\}$, $k = \{1,2,3,4,5,6\}$ (Fig. 13). Then, a new weighted eigenspace is created as follows: We first built a new eigenspace representation for the same group of people, but now using the neutral image taken at the second session (Fig. 3g). When we now test the system with the rest of the

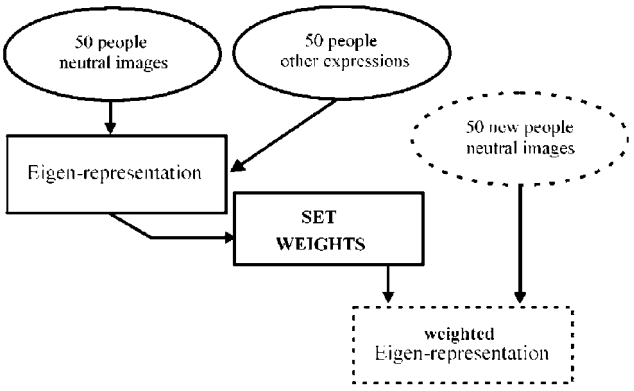


Fig. 13. Learning the weighted eigenspace representation. The 50 "new" people (dotted circle) can mean either: 1) Learning and testing do *not* use the same people or 2) both groups are the *same*, but the second eigenspace is created with different sample images than the ones used in the first step (this is done here to guarantee that the weights obtained are not optimal for the testing data).

images (Figs. 3h, 3i, and 3j), the above-computed weights are used. The recognition rates obtained are in Fig. 14a. We show in Fig. 14b results obtained with the previously unweighted version of our system (as described in Section 3 and 4). It is interesting to see that (at least for the results reported here) we were close to achieving expression-invariant recognition rates for the happy and angry faces. As expected, the "screaming" face does not approach the other two recognition rate curves, but it is important to point out the significant increase on its recognition rate.

Experiment 7. For this new test, we employed a completely new group of individuals. However, we did not recalculate the weights to make them "optimal" to this new group of people. Instead we used the weights obtained above in *Experiment 6* (with the first group of people). Due to the fact that all images of the AR-face database were taken in the same place, we assume that large cultural differences in the way individuals express emotions do not exist in this data set. The experiment works as follows. We first create the new weighted eigenspace that corresponds to the new group of people by means of the neutral expression images (Fig. 3a) of each individual of the new group and the weights obtained from the first group in the above experiment (*Experiment 6*). The images expressing happiness, anger, and "scream" (Figs. 3b, 3c, and 3d) from the new group were used for testing. The results obtained using the weighted eigenspace are shown in Fig. 14c. The results without using the weights are displayed in Fig. 14d for comparison purposes. It is the case, for our experimental data at least, that if large cultural differences do not exist between groups, significantly better results are obtained when one uses the weighted eigenspace representation.

6 Conclusions

In this paper, we have shown that it is possible for an automatic recognition system to compensate for imprecisely localized, partially occluded, and expression-variant faces even when only one single training sample per class is available. It has been shown that the localization error problem is indeed important and can sometimes lead to the incorrect classification of faces. We have solved this problem by learning the subspace within the eigenspace

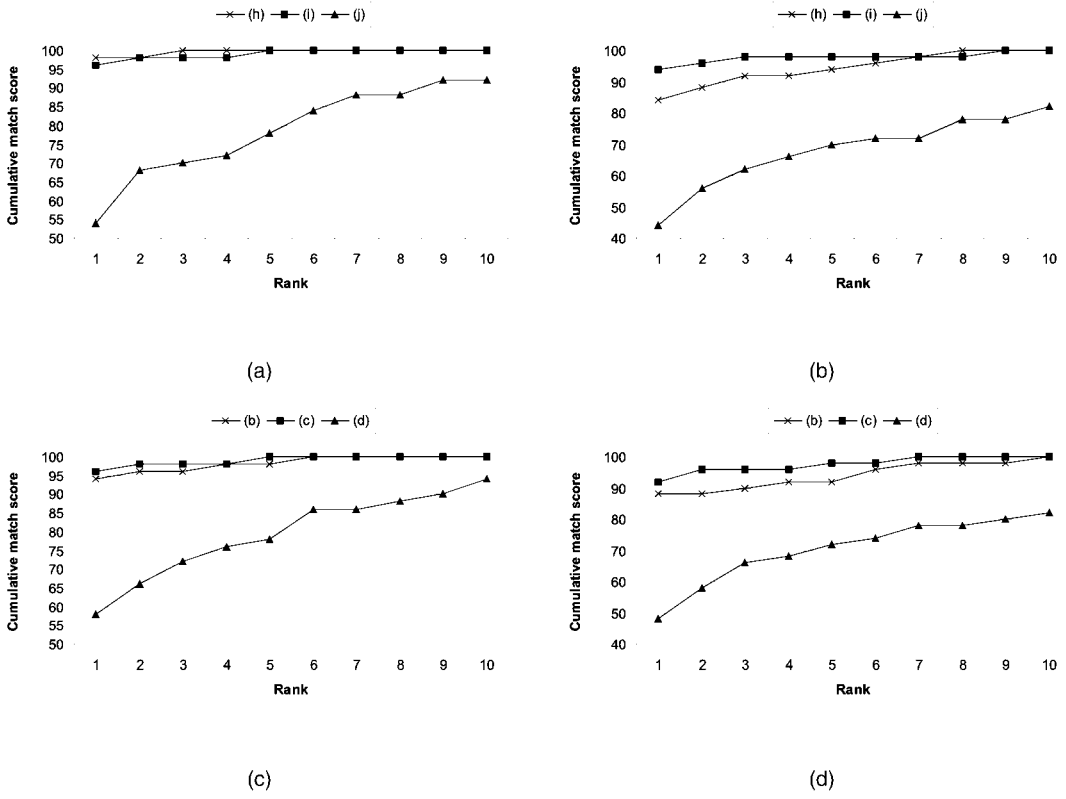


Fig. 14. Comparison of the results obtained with the weighted and unweighted eigenspaces (see text). (a) and (b) Results of *Experiment 6*. (c) and (d) Results of *Experiment 7*.

that represents this error. We report results on approximating these subspaces by means of a Gaussian and a mixture of Gaussians distribution. The latter led to slightly better results, but to an extra computational cost to be considered. A drawback of this method is that the ground-truth data (i.e., the correct localization of every feature to be localized on each face) for a set of s samples is needed in order to estimate the localization error of a given localization algorithm. The larger s is, the better. The problem with this is that the ground-truth data has to be obtained manually, which is a cost to be considered.

In order to solve the occlusion problem, a face is divided into k local parts and then a probabilistic approach is used to find the best match. The probabilities are obtained when one uses the Mahalanobis distance defined by the Gaussian distributions described above (for solving the localization problem). We demonstrated experimentally that the suppression of 1/6 of the face does not decrease accuracy. Even for those cases where 1/3 of the face is occluded, the identification results are very close to those we obtained in the nonocclusion case. We have also shown that, although intuition suggests that worse results are obtained when one occludes the eye area of a face than when one occludes the mouth area, this is not the case when duplicate images are used for testing. This leads to the conclusion that there is still much to be learned about duplicate images before we can design systems that successfully classify the images of the duplicates.

Finally, we have shown that the results of an appearance-based approach totally depend on the differences that exist between the facial expressions displayed on the learning and testing images. In an attempt to overcome this problem, we have built a weighted eigenspace representation that gives more importance to the results obtained from those local areas that are less affected by the current displayed emotion (of the testing image) and less importance to those that are more affected (from an identification point of view). A learning mechanism has been proposed and shown to be reliable where the learning and testing groups of people do not have large cultural differences that could affect the way the people in the database express certain emotions.

The importance of the probabilistic local approach is not limited to the problems reported in this paper though. It was discussed in Section 5 that the recognition of duplicates is a difficult task that must be better studied and understood. The approach proposed here could be used to study which areas of the face change less over time and why. The local approach could also be used to tackle changes due to local illumination changes. We could use systems that are robust to illumination changes and reformulate them within the local approach defined in this paper.

Acknowledgments

The author is indebted to Avi Kak, Ronnie Wilbur, Mar and Hobbes for discussion, help and proofreading. This research was partially financially supported by and partially conducted at the Sony Computer Science Lab. The author is also supported by US National Science Foundation grant no. 99-05848. The author would also like to thank the referees for their comments.

References

[1] Y. Adini, Y. Moses, and S. Ullman, "Face Recognition: The Problem of Compensating for Changes in Illumination Direction," *IEEE Trans. Pattern Analysis and Machine Intelligence,* no. 19, vol. 7, pp. 721-732, July 1997.

[2] A. Barron, J. Rissanen, and B. Yu, "The Minimum Description Length Principle in Coding and Modeling," *IEEE Trans. Information Theory,* vol. 44, no. 6, pp. 2743-2760, 1998.

[3] P.N. Belhumeour, J.P. Hespanha, and D.J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 711-720, July 1997.

[4] P.N. Belhumeour and D.J. Kriegman, "What Is the Set of Images of an Object under All Possible Lighting Conditions," *Int'l J. Computer Vision,* 1998.

[5] D. Beymer and T. Poggio, "Face Recognition from One Example View," *Science,* vol. 272, no. 5250, 1996.

[6] M.J. Black, D.J. Fleet, and Y. Yacoob, "Robustly Estimating Changes in Image Appearance," *Computer Vision and Image Understanding,* vol. 78, no. 1, pp. 8-31, 2000.

[7] R. Brunelli and T. Poggio, "Face Recognition: Features versus Templates," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 15, no. 10, pp. 1042-1053, Oct. 1993.

[8] R. Brunelli and D. Falavigna, "Person Recognition Using Multiple Cues," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 17, no. 10, pp. 955-966, Oct. 1995.

[9] R. Brunelli, "Estimation of Pose and Illumination Direction for Face Processing," *Image and Vision Computing,* vol. 15, pp. 741-748, 1997.

[10] R. Brunelli and T. Poggio, "Template Matching: Matched Spatial Filters and Beyond," *Pattern Recognition,* vol. 30, no. 5, pp. 751-768, 1997.

[11] R. Campbell, "The Lateralization of Emotion: A Critical Review," *Int'l J. Psychology,* vol. 17, pp. 211-219, 1982.

[12] H. Chen, P. Belhumeour, and D. Jacobs, "In Search of Illumination Invariants," *Proc. IEEE Computer Vision and Pattern Recognition,* vol. 1, pp. 254-261, June 2000.

[13] D.J. Clemens and D.W. Jacobs, "Space and Time Bounds on Indexing 3D Model from 2D Images," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 13, no. 10, pp. 1007-1017, Oct. 1990.

[14] I. Craw and P. Cameron, "Face Recognition by Computers," *Proc. British Machine Vision Conf.,* pp. 498-507, 1992.

[15] I. Craw, N. Costen, T. Kato, and S. Akamatsu, "How Should We Represent Faces for Automatic Recognition?" *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 21, no. 8, pp. 725-736, Aug. 1999.

[16] C. Darwin, *The Expression of the Emotions in Man and Animals.* London: John Murray, 1872. Reprinted by Univ. of Chicago Press, 1965.

[17] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Royal Statistical Soc.,* vol. 30, no. 1, pp. 1-38, 1977.

[18] G. Donato, M.S. Bartlett, J.C. Hager, P. Ekman, and T.J. Sejnowski, "Classifying Facial Actions," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 21, no. 10, pp. 974-989, Oct. 1999.

[19] T. Evgeniou, M. Pontil, and T. Poggio, "Regularization Networks and Support Vector Machines," *Advances in Computational Math.,* vol. 13, no. 1, pp. 1-50, 2000.

[20] P. Ekman and W. Friesen, *Facial Action Coding System: A Technique for the Measurements of Facial Movements.* Consulting Psychologists Press, 1978.

[21] K. Fukunaga, *Introduction to Statistical Pattern Recognition,* second ed. Academic Press, 1990.

[22] J. Hager, "Asymmetries in Facial Expression," *Emotion in the Human Face,* P. Ekman, ed. pp. 318-352, Cambridge Univ. Press, 1982.

[23] C.-Y. Huang, O.I. Camps, and T. Kanungo, "Object Recognition Using Appearance-Based Parts and Relations," *Proc. IEEE Computer Vision and Pattern Recognition,* pp. 878-884, 1997.

[24] A.K. Jain and B. Chandrasekaran, "Dimensionality and Sample Size Considerations in Pattern Recognition Practice," *Handbook of Statistics,* P.R. Krishnaiah and L.N. Kanal, eds., vol. 2, pp. 835-855, 1982.

[25] A. K. Jain, Y. Zhong, and S. Lakshmanan, "Object Matching Using Deformable Templates," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 18, no. 3, Mar. 1996.

[26] A.K. Jain, R.P.W. Duin, and J. Mao, "Statistical Pattern Recognition: A Review," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 1, pp. 4-37, Jan. 2000.

[27] R.A. Jacobs, M.I. Jordan, and G.E. Hinton, "Adaptative Mixture of Local Experts," *Neural Computer,* vol. 3, pp. 79-87, 1991.

[28] M. Kirby and L. Sirovich, "Application of the Karhunen-Love Procedure for the Characterization of Human Faces," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 12, no. 1, pp. 103-108, Jan. 1990.

[29] B. Kolb and L. Taylor, "Facial Expression, Emotion, and Hemispheric Organization," *Cognitive Neuroscience of Emotion,* R.D. Lane and L. Nadel, eds., pp. 62-83, Oxford Univ. Press, 2000.

[30] M. Lades, J.C. Vorbrüggen, J. Buhmann, J. Lange, C. von der Malsburg, R.P. Würtz, and W. Konen, "Distortion Invariant Object Recognition in the Dynamic Link Architecture," *IEEE Trans. Computers,* vol. 42, no. 3, pp. 300-311, Mar. 1993.

[31] A. Lanitis, C.J. Taylor, and T.F. Cootes, "Automatic Interpretation and Coding of Face Images Using Flexible Models," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 743-756, July 1997.

[32] M.J. Lyons, J. Budynek, and S. Akamatsu, "Automatic Classification of Single Facial Images," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 21, no. 12, pp. 1357-1362, Dec. 1999.

[33] M.J. Lyons, R. Campbell, A. Plante, M. Coleman, M. Kamachi, and S. Akamatsu, "The Noh Mask Effect: Vertical Viewpoint Dependence of Facial Expression Perception," *The Royal Soc. Proc.: Biological Sciences,* 2000.

[34] A.M. Martínez and R. Benavente, "The AR Face Database," CVC Technical Report no. 24, June 1998.

[35] A.M. Martínez and A.C. Kak, "PCA versus LDA," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 23, no. 2, pp. 228-233, Feb. 2001.

[36] A.M. Martínez, "Recognition of Partially Occluded and/or Imprecisely Localized Faces Using a Probabilistic Approach," *Proc. IEEE Computer Vision and Pattern Recognition,* vol. 1, pp. 712-717, June 2000.

[37] A.M. Martínez, "Semantic Access of Frontal Face Images: The Expression-Invariant Problem," *Proc. IEEE Content Based Access of Images and Video Libraries,* pp. 55-59, June 2000.

[38] A.M. Martínez and J. Vitrià, "Learning Mixture Models Using a Genetic Version of the EM Algorithm," *Pattern Recognition Letters,* vol. 21, pp. 759-769, 2000.

[39] G. McLachlan and K. Basford, *Mixture Models: Inference and Applications to Clustering.* Marcel Dekker, 1988.

[40] G. McLachlan and T. Krishnan, *The EM Algorithm and Extensions.* Wiley, 1997.

[41] S.K. Mishra and V.V. Raghavan, "An Empirical Study of the Performance of Heuristic Methods for Clustering," *Pattern Recognition in Practice,* E.S. Gelsema and L.N. Kanal, eds., pp. 425-436, North Holland, 1994.

[42] T.M. Mitchell, *Machine Learning.* McGraw-Hill, 1997.

[43] B. Moghaddam and A. Pentland, "Probabilistic Visual Learning for Object Representation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 696-710, July 1997.

[44] B. Moghaddam and A. Pentland, "Beyond Euclidean Eigenspaces: Probabilistic Matching for Visual Recognition," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* 1998.

[45] M. Moscovitch and J. Olds, "Asymmetries in Spontaneous Facial Expressions and Their Possible Relation to Their Hemispheric Specialization," *Neuropsychologia,* vol. 20, pp. 71-81, 1982.

[46] K. Ohba and K. Ikeuchi, "Detectability, Uniqueness, and Reliability of Eigen Windows for Stable Verification of Partially Occluded Objects," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 9, pp. 1043-1048, Sept. 1996.

[47] M. Pantic and L.J.M. Rothkrantz, "Automatic Analysis of Facial Expressions: The State-of-the-Art," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 12, pp. 1424-1445, Dec. 2000.

[48] P.S. Penev and J.J. Atick, "Local Feature Analysis: A General Statistical Theory for Object Representation," *Network: Computation in Neural Systems,* vol. 7, no. 3, pp. 477-500, 1996.

[49] P.S. Penev and L. Sirovich, "The Global Dimensionality of Face Space," *Proc. IEEE Face and Gesture Recognition,* pp. 264-270, 2000.

[50] A. Pentland, T. Starner, N. Etcoff, N. Masoiu, O. Oliyide, and M. Turk, "Experiments with Eigenfaces," *Proc. Workshop Int'l Joint Conf. Artificial Intelligence, Looking at People,* 1993.

[51] P.J. Phillips, H. Moon, P. Rauss, and S.A. Rizvi, "The FERET Evaluation Methodology for Face-Recognition Algorithms," *Proc. First Int'l Conf. Audio and Video-Based Biometric Person Authentification,* 1997.

[52] S. Richardson and P. Green, "On Bayesian Analysis of Mixtures with Unknown Numbers of Components," *J. Royal Statistics Soc. B,* vol. 59, pp. 731-792, 1997.

[53] K. Rose, "Deterministic Annealing for Clustering, Compression, Classification, Regression and Related Optimization Problems," *Proc. IEEE,* vol. 86, pp. 2210-2239, 1998.

[54] H.A. Rowley, S. Baluja, and T. Kanade, "Neural Network-Based Face Detection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 1, pp. 23-38, Jan. 1998.

[55] H. Sackheim, R.C. Gur, and M.C. Saucy, "Emotions Are Expressed More Intensively on the Left Side of the Face," *Science,* vol. 202, pp. 434-436, 1978.

[56] P. Simard, Y. LeCun, and J. Denker, "Efficient Pattern Recognition Using a New Transformation Distance," *Advances in Neural Information Processing Systems 5,* S.J. Hanson, J.D. Cowan, and C.L. Giles, eds., Morgan Kaufmann, 1993.

[57] P. Simard, B. Victorri, Y. LeCun, and J. Denker, "Tangent Prop–A Formalism for Specifying Selected Invariances in an Adaptative Network," *Advances in Neural Information Processing Systems 4,* J.E. Moody, S.J. Hanson, and R.P. Lippmann, eds. pp. 651-655, Morgan Kaufmann, 1992.

[58] L. Sirovich and M. Kirby, "Low-Dimensional Procedure for the Characterization of Human Faces," *J. Optical Soc. Am. A,* vol. 4, pp. 519-524, 1987.

[59] K.-K. Sung and T. Poggio, "Example-Based Learning for View-Based Human Face Detection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 1, pp. 39-51, Jan. 1998.

[60] M.J. Swain and D.H. Ballard, "Color Indexing," *Int'l J. Computer Vision,* vol. 7, no. 1, pp. 11-32, 1991.

[61] Y.-I. Tian, T. Kanade, and J.F. Cohn, "Recognizing Action Units for Facial Expression Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 23, no. 2, pp. 97-115, Feb. 2001.

[62] G.V. Trunk, "A Problem of Dimensionality: A Simple Example," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 1, no. 3, pp. 306-307, 1979.

[63] M. Turk and A. Pentland, "Eigenfaces for Recognition," *J. Cognitive Neuroscience,* vol. 3, no. 1, pp. 71-86, 1991.

[64] S. Ullman, *High-Level Vision: Object Recognition and Visual Cognition.* MIT Press, 1996.

[65] V.N. Vapnik, *Statistical Learning Theory.* New York: Wiley, 1998.

[66] T. Vetter and T. Poggio, "Linear Object Class and Image Synthesis from a Single Example View," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 733-742, July 1997.

[67] T. Vetter and N.F. Troje, "Separation of Texture and Shape in Images of Faces for Image Coding and Synthesis," *J. Optical Soc. Am. A,* vol. 14, no. 9, pp. 2152-2161, 1997.

[68] M. Whindham and A. Cutler, "Information Ratios for Validating Mixture Analysis," *Int'l J. Computer Vision,* vol. 28, no. 2, pp. 103-116, 1998.

[69] L. Wiskott and C. von der Malsburg, "Recognizing Faces by Dynamic Link Matching," *Proc. Int'l Conf. Artificial Neural Networks,* pp. 347-352, 1995.

[70] L. Wiskott, J. Fellous, N. Krüger, and C. von der Malsburg, "Face Recognition by Elastic Bunch Graph Matching," *Intelligent Biometric Techniques in Fingerprint and Face Recognition,* L.C. Jain et al. eds., Springer-Verlag, 1999.

[71] J. Yang and A. Waibel, "A Real Time Face Tracker," *Proc. Third IEEE Workshop Applications of Computer Vision,* 1996.

[72] M.-H. Yang, D.J. Kriegman, and N. Ahuja, "Detecting Faces in Images: A Survey," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 1, pp. 34-58, Jan. 2002.

[73] A. Yuille, P. Hallinan, and D. Cohen, "Feature Extraction from Faces Using Deformable Templates," *Int'l J. Computer Vision,* vol. 8, no. 2, pp. 99-111, 1992.

[74] W. Zhao and R. Chellapa, "SFS Based View Synthesis for Robust Face Recognition," *Proc. IEEE Face and Gesture Recognition,* pp. 285-292, 2000.

Aleix M. Martínez is a research associate in the Electrical and Computer Engineering Department at Purdue University. He has also been affiliated with the Sony Computer Science Lab. He has worked on projects related to computer vision, computational neuroscience, robotics, autonomous agents, and computational linguistics and is the author of several papers along these lines. He is the author (with R. Benavente) of the AR-face database. He serves as a referee for several journals and is on the program committee of several conferences. He is a member of the IEEE.