Available online at www.sciencedirect.com
SCIENCE DIRECT
Image and Vision Computing 24 (2006) 626–638
image AND vision COMPUTING
www.elsevier.com/locate/imavis

# A weighted probabilistic approach to face recognition from multiple images and video sequences

Yongbin Zhang, Aleix M. Martínez *

*Department of Electrical and Computer Engineering, The Ohio State University, 2015 Neil Ave, Columbus, OH 43210, USA*

Received 29 November 2004; received in revised form 27 July 2005; accepted 23 August 2005

**Abstract**

To date, advances in face recognition have been dominated by the design of algorithms that do recognition from a single test image. Recently, an obvious but important question has been put forward. Will the recognition results of such approaches be generally improved when using multiple images or video sequences? To test this, we extend the formulation of a probabilistic appearance-based face recognition approach (which was originally defined to do recognition from a single still) to work with multiple images and video sequences. In our algorithm, as it is the case in most appearance-based approaches, we will need to use a feature extraction algorithm to find those features that best describe and discriminate among face images of distinct people. We will show that regardless of the algorithm used, the recognition results improve considerably when one uses a video sequence rather than a single still. Hence, a positive answer to our question (in the general sense) seems reasonable. The probabilistic algorithm we propose in this paper is robust to partial occlusions, orientation and expression changes, and does not require of a precise localization of the face or facial features. We will also show how these problems are more easily solved when one uses a video sequence rather than a single image for testing. The limitations of our algorithm will also be discussed. Understanding the limitations of current techniques when applied to video is important, because it helps identify those weak points that require further consideration.


## 1. Introduction

As computers become more ubiquitous and technology more accessible to the end user, face recognition systems are expected to play an increasingly important role in society. A recent change that has influenced the way we study object recognition and other computer vision problems is the improvement in quality and reduction in price of video cameras. This has drawn interests in the design of systems that can recognize objects (such as faces) from video sequences rather than from stills.

Compared to the number of algorithms that do recognition from stills, the literature on video-based (or multi-image-based) methods is relatively small [43]. One reason for this imbalance was due to the low accessibility of high-quality video cameras. The second reason is algorithmical. While it is generally difficult to successfully do feature extraction from still images, this process has proven even more challenging over dynamic sequences [23,9,26]. This second point raises an important question. Would the methods, defined to recognize faces from a single test image, perform better if they could work with multiple images or video sequences? Note that if the answer to this question were positive, there would be less need for the design of feature extraction algorithms that can do a more direct analysis of dynamic sequences. Understanding the limitations of current algorithms when applied to video will help researchers design algorithms that can specifically solve these problems [34,16].

To answer our question though, we need to be able to re-define our previous approaches (originally intended to work with stills) so that they are capable of combining the information extracted from multiple frames. In this paper, we first present an extension of our probabilistic approach defined in [27] to one that can properly combine the information extracted from multiple frames. We then introduce an algorithm that selects those images of the test video sequence that are most useful for recognition. Finally, we show that when combining these two ideas, the classification results can increase considerably.

The learning algorithm is defined to work with multiple training images too. These may be part of a video sequence, correspond to images captured with distinct cameras, from different orientations, etc. At this stage, we will also use a linear feature extraction method to find that feature space that

* Corresponding author.
*E-mail address:* aleix@ece.osu.edu (A.M. Martínez).

0262-8856/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2005.08.004

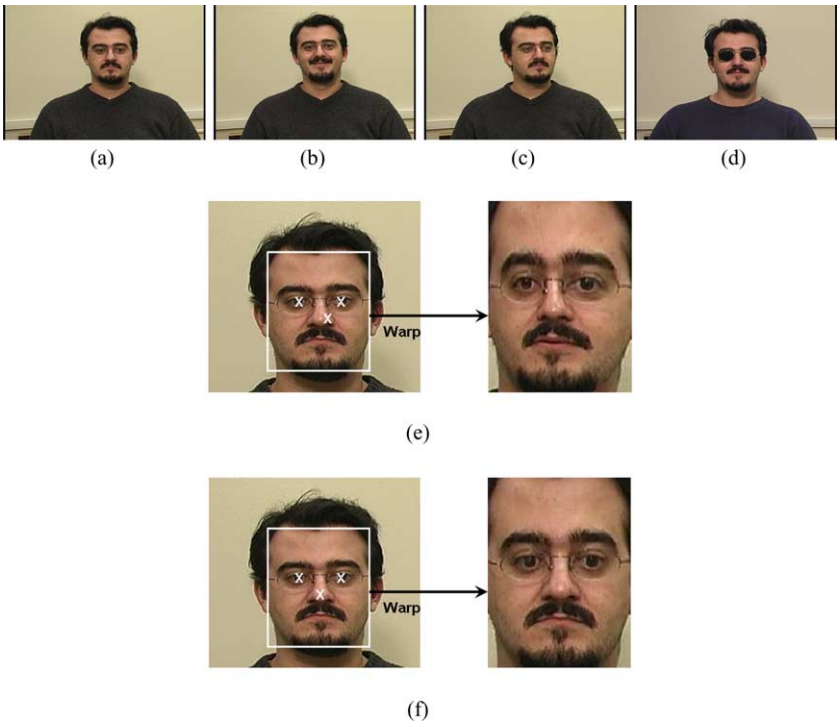


Fig. 1. (a)–(d) Shown here are four images of the same person but with different facial expression (a) and (b), pose (a) and (c), and with an occlusion (d). (e) Shows the effects caused by imprecisely localized features on the warped image as compared to the correct result, which is illustrated in (f).

best discriminates among classes (i.e. people). In this paper, we test three of the most commonly used techniques: principal component analysis (PCA), independent component analysis (ICA), and linear discriminant analysis (LDA). We will show that indeed the use of video sequences can boost the results of most known algorithms regardless of the (linear) feature extraction method used.

The video-based method presented in this contribution is robust to occlusions, expression and pose changes and errors of localization; see Fig. 1(a)–(d). To make the system robust to occlusions, we will divide each image frame into a set of local areas, each of which will be analyzed independently. The recognition results in each of the local areas (for all of the images in the video sequences) will be combined using a probabilistic framework. To address the problems of expression and pose changes, we will find those images and image features of the test video sequence that are more alike to the ones used for training. The most similar ones will be assigned a higher weight, the most dissimilar a lower weight.

Another problem in face recognition is that caused by errors of localization. This problem is exacerbated when one works with video sequences. Imprecise localizations are given whenever a face or facial feature (that needs to be automatically localized in the image) is not detected with pixel precision (i.e. the position of the actual features and those estimated by the algorithm are not the same but close) [27]. This can produce a feature vector that is closer to the subset of an incorrect class. For example, in the appearance-based framework, one needs to first warp the face to a standard shape before proper comparisons can be carried out [5,40]. In these cases, imprecise localizations will result in a shape change that will be seen as a distracter by our classifier; see Fig. 1(e) and (f). To solve this, the subset of all images under all possible errors of localization needs to be estimated. We will show how we can model this and use such estimates to define the probabilities of each of the classes.

The rest of this paper is organized as follows. In Section 2, we compare some of the previous work presented in the literature to that presented in this paper. Sections 3 and 4 detail on the formulation of our approach. Experimental results are in Section 5. Conclusions are in Section 6.

## 2. Previous work

In the last few years, several algorithms that attempt to do recognition using video sequences have been proposed. The goal is to find an image representation, a feature extractor and a classifier that allows to improve upon the results generated by those algorithms that use a single still. In 2002, this goal seemed yet more challenging, when the results of the 2002 face recognition vendor test could not find any clear advantage (i.e. higher recognition rates) when using video sequences rather than stills [33]. Since then, interest in the design of robust algorithms has grown. However, the methods reported so far do not simultaneously address the problems posed by occlusion, expression and pose changes, and errors of localization.

Gong et al. [15] used a feed-forward neural network approach to do recognition of faces from video sequences. In their paper, the weights of the network are associated to each

of the testing images. These are then modified so as to maximize classification accuracy. In [19], the authors use radial basis function (RBF) networks instead. In this case, the learning algorithm uses multiple stills that are captured from different orientations. Recognition is then done over a video sequence. The algorithm is thus robust to pose variations but not to occlusions or expression changes. A similar algorithm was also defined by Wechsler et al. in Ref. [41].

Another approach defined by Li and Chellappa [25] attempts to classify faces according to the motion parameters of a set of facial features that are tracked over time. This method can be made robust to different orientations (if the pose is known) and to occlusions (if the locations of such occlusions are also known). However, the method is sensitive to facial expression changes and errors of localization.

Edwards et al. [10] use the active appearance model [7] to estimate the variability of the face within a video sequence. This can then be used to improve learning and identification. Hidden Markov models can also be used to learn the appearance changes from a sequence of images [26]. These methods, however, require of a large number of samples to successfully extract the most discriminant information of each individual. In [24], the authors propose to learn view-based manifolds from video. Robustness is increased with the help of a transition probability matrix that is defined between adjacent views. This can be used to define algorithms invariant to pose.

Zhou and Chellappa [44] define a recognition system that uses sequential importance sampling to approximate the joint probability distribution of identity and motion. Their framework may be used to define algorithms robust to pose and expression. And, finally, [35] addresses the problem of recognizing faces from images acquired over long-term observations. Classification is given by the relative entropy between the probability density functions of the testing and training set.

The systems summarized above do not however address the problems of errors of localization, occlusions, and expression changes and pose variation simultaneously. The question remains: is it possible to define a video-based recognition algorithm that properly addresses these problems and achieves higher recognition rates than those obtained using a single image for testing? In this paper, we will show that the answer to this question is indeed positive.

Our first step is to extend our probabilistic approach, originally presented in Ref. [27] (for a recent survey, see [29]), to work with video sequences rather than stills. We will then show how we can extend our formulation to make the approach more robust to expression changes and pose variations. Here, we will search for those images (of our test sequence) where the facial expression and pose are most similar to those seen in the training samples. We present experimental results on a recently created database and show that our method is superior to that defined in Ref. [27] and to those algorithms based on PCA, ICA and LDA [36,39,1,11,3].

## 3. Imprecisely localized and partially occluded faces

### 3.1. Modelling the subset of localization error

Any face recognition system requires of a pre-localization of the face and facial features. The localization of these facial features is necessary to either construct feature-based representations of the face or to warp all faces to a standardized size (and/or shape) for appearance-based recognition [5,17,27,43]. However, it would be unrealistic to hope for any such system to be able to localize every facial feature (e.g. eyes, mouth, etc.) with pixel precision. 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 result in an incorrect classification. In Ref. [27], we proposed to model the subset within the feature space that contains most of the images generated under all possible localizations (including all possible small errors of localization). An extension of this approach is presented below.

For each of the training images, the goal is to find that subset (within the feature-space) that represents all images under all possible errors of localization. We estimate this subset using a mixture of multivariate Gaussian distributions. More formally, let $\{\mathbf{I}_{1,1},\ldots,\mathbf{I}_{1,n_1},\ldots,\mathbf{I}_{C,n_C}\}$ be the set of training images, $C$ the number of class, and $n_i$ the number of samples in class $i$. Since, the average localization error of our localization algorithm is known or can be estimated using a labelled set, one can synthetically generate the set of all images under all possible errors of localization for each of the training samples; i.e. $\hat{I}_{i,j}=\{\mathbf{I}_{i,j,1},\ldots,\mathbf{I}_{i,j,r}\}$, where $i$ is the class, $j$ specifies the sample number of class $i$ and $r$ is the maximum number of images we can generate with the known error [27][1]. Once the dataset $\hat{I}_{i,j}$ has been generated, the subset where all these possible warped faces lay, can be modelled using an appropriate probability density function. Since the subset $\hat{I}_{i,j}$ may be non-linear, a mixture of Gaussians is a good choice. Mixture of Gaussians can be estimated using the expectation–maximization (EM) algorithm [8], which is an iterative method divided into two steps, the E-step

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

and the M-step

$$\begin{aligned}\mu_{i,j,g}^{[t+1]}&=\frac{\sum_{l=1}^{r}h_{i,j,l,g}^{[t]}\mathbf{I}_{i,j,l}}{\sum_{l=1}^{r}h_{i,j,l,g}^{[t]}},\\ \sum\nolimits_{i,j,g}^{[t+1]}&=\frac{\sum_{l=1}^{r}h_{i,j,l,g}^{[t]}(\mathbf{I}_{i,j,l}-\mu_{i,j,g}^{[t+1]})(\mathbf{I}_{i,j,l}-\mu_{i,j,g}^{[t+1]})}{\sum_{l=1}^{r}h_{i,j,l,g}^{[t]}},\end{aligned} \tag{2}$$

[1] Since these images will usually be highly correlated, a subset of all these should suffice in practise.

where $\mu_{i,j,g}$ and $\Sigma_{i,j,g}$ are the $g$th mean and covariance matrix of the $j$th sample in class $i$, $G$ is the total number of models used in the mixture (i.e. number of Gaussians) and $[t]$ means iteration $t$. And, in our formulation, it is assumed that all models (Gaussians) are equally probable. The original estimates for each of the means and covariance matrices are obtained by small random modifications of the sample mean and covariance matrix of each class.

### 3.2. Dealing with occlusions

In order to be robust to partial occlusions, we divide the face into $K$ (ellipsoidal-shaped) local areas, as shown in Fig. 2. Note that each sample image $\mathbf{I}_{i,j}$ will generate $r$ possible images (to account for the localization error) for each of the $K$ subimages. Therefore, the new set of images will be given by $\{\mathbf{I}_{i,j,1,1}, \ldots \mathbf{I}_{i,j,1,r}, \ldots \mathbf{I}_{i,j,K,r}\}$, where $\mathbf{I}_{i,j,k,l}$ is the $l$th sample (accounting for the localization error) of the $k$th local area of the $j$th sample in class $i$.

As above, we estimate the subset of the localization error of every local region by means of a mixture of $G$ Gaussians, i.e. $\{\mu_{i,j,k,g}\}_{g=1}^{G}$ are the $G$ sample means of the $G$ Gaussians for each of the $K$ subareas of the $j$th sample in class $i$, and $\{\sum_{i,j,k,g}\}_{g=1}^{G}$ the corresponding sample covariance matrices. Fig. 2 shows the subsets representing all images under all possible errors of localization for one of the local areas of two images (each image of a different person). In the figure, $G=3$.

### 3.3. Subspace representation

In appearance-based algorithms, the original space corresponds to a dense pixel representation. This generally means that its dimensionality is too large to allow the computation of the subset of the localization error from it, because the number of samples is smaller than the number of features (dimensions). In such cases, it is convenient to reduce the dimensionality of our original feature space to one where the number of samples per dimension is appropriate. Many subspace techniques have been used to achieve this among which principal component analysis (PCA) [13,36], linear discriminant analysis (LDA) [12,13] and independent component analysis (ICA) [20] have arguably been the most popular. Each subspace technique will generate a distinct projection matrix. In the rest of this paper, we will use $\Phi_a$ to refer to the projection matrix obtained with method $a$; e.g. $\Phi_{\mathrm{PCA}}$ is the projection matrix obtained using the PCA approach.

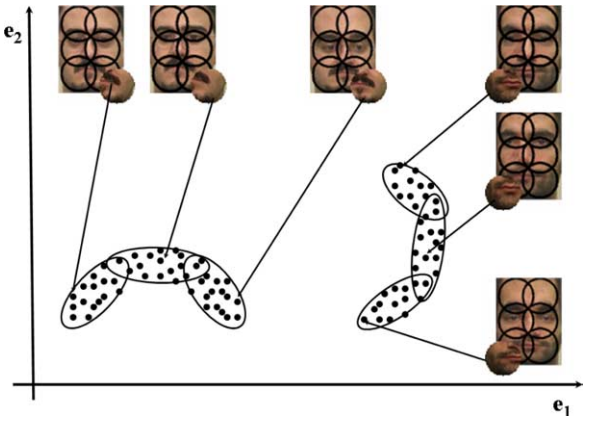


Fig. 2. We generate all possible warped face images according to the localization error of our localization algorithm. Then, each face image is divided into $K$ local areas. The localization error is estimated (for each of these local areas) using a mixture of Gaussians.

The columns of $\Phi_{\mathrm{PCA}}$ are the $e$ eigenvectors associated to the $e$ largest eigenvalues of the sample covariance matrix of the data $\Sigma_X$. This guarantees that we will minimize the reconstruction error of the Gaussian distribution, which is given by the central moments of the data $\Sigma_X$. If the discriminant structure is however hidden in moments of higher order, PCA may not obtain the most discriminant ($e$-dimensional) subspace. Unlike PCA, ICA uses higher moments of the data to find those feature vectors that are most independent from each other [20,1]. Unfortunately, ICA does not have a general close-form solution, and iterative methods need to be used instead. In our experimental results, we use the Infomax algorithm defined in [4].

While PCA and ICA are unsupervised algorithms, LDA works with labelled (training) data. This facilitates the selection of those basis vectors that maximize the distance between the means of each class and minimize the distance between the samples in each class and their corresponding class means [12,13].

Since, we have divided the face into $K$ sub-blocks, we will need to generate $K$ projection matrices $\Phi_{a,k}$, where $k=\{1, \ldots, K\}$ and (as above) $a=\{\mathrm{PCA}, \mathrm{ICA}, \mathrm{LDA}\}$. Once the set of these projection matrices $\{\Phi_{a,k}\}_{k=1}^{K}$ has been obtained, the subset that describes the localization errors of each sample image can be represented as $\{\hat{\mu}_{i,j,k,g}, \hat{\sum}_{i,j,k,g}\}_{g=1}^{G}$, where $\hat{\mu}_{i,j,k,g} = \Phi_{a,k}^{T}\mu_{i,j,k,g}$ and $\hat{\sum}_{i,j,k,g} = \Phi_{a,k}\sum_{i,j,k,g}\Phi_{a,k}$.

### 3.4. Identification from multiple images

Given a set of test images $\mathbf{T}=\{\mathbf{T}_1, \ldots, \mathbf{T}_m\}$, we work as follows. First, we divide each image $\mathbf{T}_p$ into the $K$ local areas defined above, $\{\mathbf{T}_{p,1}, \ldots, \mathbf{T}_{p,K}\}$. Then, the probability of each class $c$ in each local area can be estimated as

$$P(c|\mathbf{T}_{p,k}) = \max_{j,g} (\hat{T}_{p,k} - \hat{\mu}_{c,j,k,g})^{\mathrm{T}} \hat{\sum}_{c,j,k,g} (\hat{T}_{p,k} - \hat{\mu}_{c,j,k,g}), \quad (3)$$

where $\hat{T}_{p,k} = \Phi_{a,k}^{T}\mathbf{T}_{p,k}$. The probability of class $c$ given the observation vector $\mathbf{T}_p$ is

$$P(c|\mathbf{T}_p) = \frac{\sum_{k=1}^{k} P(c|\mathbf{T}_{p,k})}{\sum_{c=1}^{C}\sum_{k=1}^{K} P(c|\mathbf{T}_{p,k})}. \quad (4)$$

The denominator in Eq. (4) is necessary to guarantee these probabilities take value in the intervale [0,1].

Ideally, we want to find the class that maximizes the total probability; i.e.

$$\arg\max_{c} \frac{1}{m}\sum_{p=1}^{m} P(c|\mathbf{T}_p). \quad (5)$$

However, in general, it is not necessary to compute the probability for all images. Usually, after using a few of them, the probability of correct match for one of the classes is much higher than the rest. To guarantee success in such cases tough, we may

want to impose an additional constraint: that the probability of the class selected be above a predefined threshold.

## 4. A Weighted probabilistic approach for video sequences

We propose to assign higher weights to those images of the test video sequence and those local areas for which the expression and viewing angle are similar to those in the training set.

### 4.1. Facial expression changes

As mentioned above, the facial expressions in the training set can diverge from those in the test sequence. This can cause problems, since it is known that the misclassifications error rate increases when the expressions in the training and testing set differ [28]. Ideally, we would like to use those images of the sequence that have a similar expression to those in the training set while discarding the rest.

To solve this problem, we first need to estimate the difference in expression between the test images and each of the training samples. A common way to do this is by means of the optical flow approach [5,2,42,27]. More formally, we write $\mathbf{O}(\mathbf{T}_p, \mathbf{I}_j)$ to represent the optical flow vector that describes the motion necessary to move the facial muscles from the expression in $\mathbf{T}_p$ to that of $\mathbf{I}_j$, where here, $\mathbf{I}_j$ is the $j$th sample image.

Once we have calculated $\mathbf{O}(\mathbf{T}_p, \mathbf{I}_j)$, we can estimate the 'usefulness' of each image in our video sequence as $W_{p,j} = O_{max} = \|\mathbf{O}(\mathbf{T}_p, \mathbf{I}_j)\|$, where $O_{\max} = \max_{\mathbf{T}_p,\mathbf{I}_j} \|\mathbf{O}(\mathbf{T}_p, \mathbf{I}_j)\|$ and $\|\cdot\|$ is the sum of the (2-norm) magnitude of the flow. Small values in $W_{p,j}$ means that the expressions in each of the two images ($\mathbf{T}_p$ and $\mathbf{I}_j$) are very different. Large values imply similar expressions in both images.

We can now use these values to weight each of the probabilities in Eq. (3). To do this, we will need to calculate the usefulness of each of the local areas, which is given by

$$W_{p,j,k} = O_{k_{max}} - \|\mathbf{O}(\mathbf{T}_{p,k}, \mathbf{I}_{j,k})\| \tag{6}$$

where $O_{k_{\max}} = \max_{\mathbf{T}_{p,k},\mathbf{I}_{j,k}} \|\mathbf{O}(\mathbf{T}_{p,k}, \mathbf{I}_{j,k})\|$, with obvious notation. We then normalize these values appropriately,

$$\hat{W}_{p,j,k} = \frac{W_{p,j,k}}{\sum_{j=1}^{n} W_{p,j,k}}, \tag{7}$$

where $n = \sum_{c=1}^{C} n_c$. We now redefine Eq. (4) as

$$P(c|\mathbf{T}_p) = \frac{\sum_{k=1}^{K} \hat{W}_{p,c,k} P(c|\mathbf{T}_{p,k})}{\sum_{c=1}^{C}\sum_{k=1}^{K} \hat{W}_{p,c,k} P(c|\mathbf{T}_{p,k})}. \tag{8}$$

The reader may have noted that since, the motion field between each pair of training and testing images is now known (i.e. pixel correspondences are given), we could have morphed one of the faces to equal the shape of the other. We have indeed experimented with this idea and observed that the results were comparable to those obtained with the approach described above but with a higher computational cost [29].

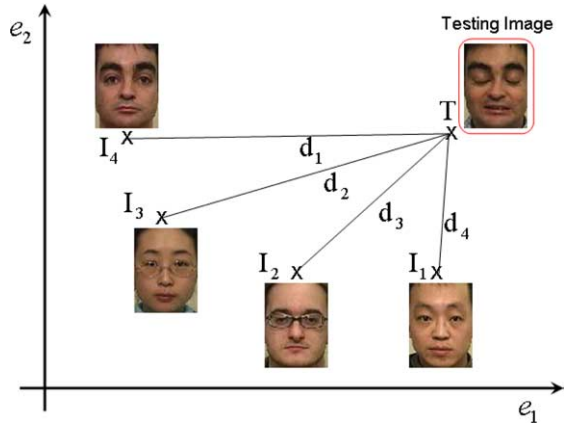


Fig. 3. This figure shows the problem posed by subspace methods when training and testing images differ in expression. In these cases, Euclidean distances may fail to classify the faces correctly.

### 4.2. Pose variations

When the face changes orientation with respect to a fix camera, the observed textural patterns in the image vary. This problem is emphasized when one uses appearance-based approaches. While in shape-based algorithms one can recover the frontal view from an estimation of the object's three-dimensional shape, in appearance-based this can only be achieved if the shape and radiometric information are properly obtained. This makes the problem of recovering the frontal view of an arbitrarily illuminated face extremely hard.

Similarly to what we did in our previous section, we can use those images of the video sequence that have a similar pose to those images in the training set and discard the rest. To achieve this, we work as follows. We first estimate the pose (with respect to the camera) of the training and testing images. Then, we weight each of the testing images according to the similarity of their pose with those observed in the training set. Identical pose in both images will require a weight of one, while pose differences of over 90 degrees should correspond to zero. All other images will be weighted inverse proportionally to their pose difference (Fig. 3).

The view-based approach advanced by Pentland and colleagues [31] has been extensively used for object recognition under varying pose as well as for pose estimation [30,37]. To do pose estimation, we first need to learn the subspace (e.g. eigenspace) that represents the images of all faces as viewed from a given orientation $\alpha$[2]. Formally, let $\Sigma_\alpha$ be the sample covariance matrix of a set of sample images with faces at orientation $\alpha$. Then, the eigenspace representing those face images at orientation $\alpha$ is given by the projection matrix $E_\alpha$ whose column vectors are the eigenvectors associated to the $r$ largest eigenvalues of $\Sigma_\alpha$. If our sample images correspond to faces photographed at $m$ distinct orientations, we will generate $m$ eigenspaces; i.e. $\{\mathbf{E}_{\alpha_1}, \ldots, \mathbf{E}_{\alpha_m}\}$. Fig. 4 illustrates this.

[2] In this paper, we have only considered rotations about the $y$-axis, but our formulation can be extended to deal with the more general (although rare) case where all three angles are considered.

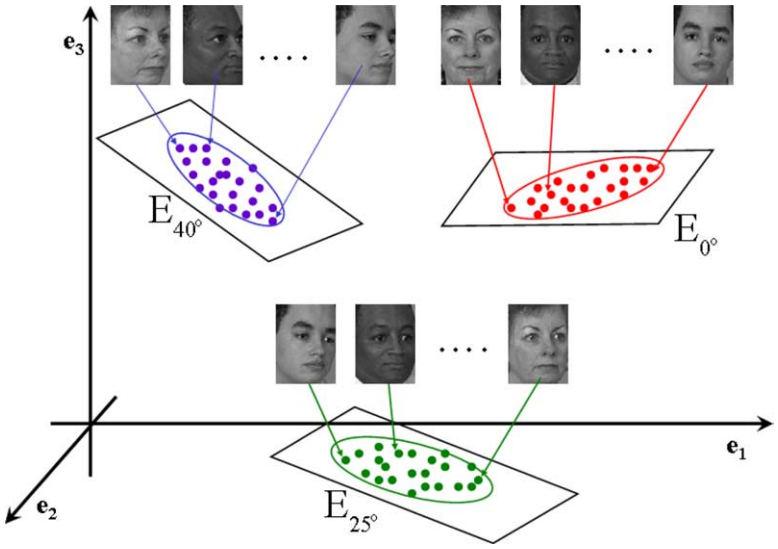


Fig. 4. Each of the subspaces (and subsets) shown here corresponds to the space (and set) representing all faces seen from a common orientation.

We can now estimate the pose of a new test image $\mathbf{T}_p$ as follows. First, we calculate the distance from the vector $\mathbf{T}_p$ to each of the subspaces. This is usually estimated with the Euclidean distance from $\mathbf{T}_p$ to the eigenspace and the Mahalanobis distance within this [38]. We will denote these distances as $\{D_{\alpha_1}, \ldots D_{\alpha_m}\}$. Next, we select the $d$ smallest distances (with $d \leq m$) and estimate the pose of $\mathbf{T}_p$ as

$$\theta_{\mathbf{T}_p} = \sum_{i=1}^{d} \frac{D_{\alpha(d-i+1)}}{\sum_{j=1}^{d} D_{\alpha_j}} \alpha_i \tag{9}$$

where $\alpha_i$ corresponds to the pose of the face images represented by the $i$th closest eigenspace $\mathbf{E}_{\alpha_i}$.

To create the view-based eigenspace defined in the preceding paragraphs, we have used the images of the FERET database [32]. This database contains a set of images taken at different poses for a total of about 200 people. This set contains images as seen from the following viewing angle: $-60°$, $-40°$, $-25°$, $-15°$, $0°$, $60°$, $40°$, $25°$ and $15°$ (where $0°$ represent the frontal view). This means, we have constructed nine eigenspaces (i.e. $m=9$), one for each orientation. We have tested this representation using the images of the FERET dataset with the leave-one-pose-out test. In this test, we leave the images corresponding to one of the nine poses out (excluding those seen from $-60°$ and $60°$, because these cannot be interpolated with Eq. (9)) and, then, use the images left out for testing. The mean error and standard deviation were $5.59°$ and $1.26°$, respectively.

Once the pose for each pair of training and testing images has been computed, we can calculate the weights. As we argued above, the weights $W'_{p,c,j}$ should be equal to one when the orientation in both images is identical, and zero when the difference is equal to or larger than 90 degrees; i.e. $W'_{p,c,j}=1$ when $\theta_{T_p} = \theta_{I_{c,j}}$ and $W'_{p,c,j}=0$ when $|\theta_{\mathbf{T}_p} - \theta_{\mathbf{I}_{c,j}}| \geq 90°$. Any other difference will be approximated using the following function

$$W'_{p,c,j} = 1 - \frac{g(|\theta_{\mathbf{T}_p} - \theta_{\mathbf{I}_{c,j}}|)}{90} \tag{10}$$

where

$$g(\theta) = \begin{cases} \theta, & \text{if } \theta \leq 90 \\ 90, & \text{otherwise.} \end{cases}$$

We finally redefine Eq. (8) as

$$P(c|\mathbf{T}_p) = \frac{\sum_{k=1}^{K} W'_{p,c,k} \hat{W}_{p,c,k} P(c|\mathbf{T}_{p,k})}{\sum_{c=1}^{C} \sum_{k=1}^{K} W'_{p,c,k} \hat{W}_{p,c,k} P(c|\mathbf{T}_{p,k})}. \tag{11}$$

where $W'_{p,c,k}$ measures the similarity in pose between the test image $\mathbf{T}_p$ and the training images $\mathbf{I}_{c,j}$ with $j$ being selected by Eq. (3).

### 4.3. Rejecting outliers from the image sequence

Although our algorithm was defined to be robust to small errors of localization, this (or any other) algorithm will be unable to recognize faces where the localization error is large. Note that, in these cases, the warped image may not even be a face. Hence, it is necessary to define a method that can select those images of the video sequence that have been localized within a small margin of error. To achieve this, we will use techniques defined in the area of robust statistics.

Our method is based on the following assumption: the number of images where the localization error is manageable (small) corresponds to more than 50% of the images in the video sequence. With this in mind, we can construct a feature space that represents the location of each of the detected facial features, $\vec{a} = (a_1, a_2, \cdots, a_h)^T$; for example, $a_1$ and $a_2$ may represent the $y$-coordinate position of the center of both eyes and $a_3$ the $x$-coordinate of the nose. Thus, each test image $\mathbf{T}_p$ will give a feature vector $\vec{a}$; where $\vec{a} \in \Re^h$ and $h$ is the number

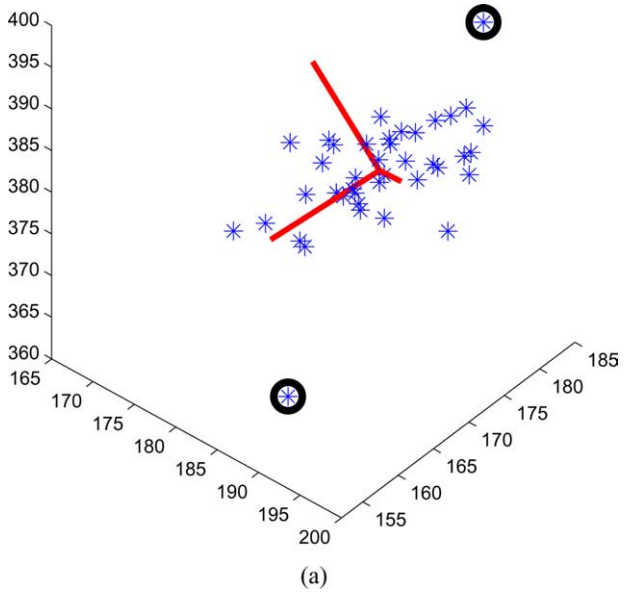


Fig. 5. To detect outliers, we define a robustified version of the covariance matrix using the median absolute deviation criterion. To illustrate how our method works, we show the results of our algorithm as it applies to a simple case. The two samples marked with a circle are the outliers automatically detected by our method.

of features needed to warp the face to its standard shape. Since the classical sample covariance matrix does not represent the data well in the presence of outliers, we calculate the covariance matrix of all those feature vectors using a robust equation [22]. Here, the covariance matrix is built with those orthonormal vectors $\{\mathbf{b}_1^T, \cdots, \mathbf{b}_q^T\}$ that maximize a given criterion $S$ when this is applied to $\mathbf{b}_i\mathbf{X}$, where each row of the matrix $\mathbf{X}$ corresponds to a feature vector for each of the test images. Gather et al. [14] showed that we can compute a robust covariance matrix by using the median absolute deviation (MAD) as a criterion for $S$.

Those feature vectors that are not represented by the robust covariance matrix described above are then considered outliers and, hence, eliminated from consecutive computations. This means that all those test images that correspond to such outliers will not be used by our algorithms; i.e. will not be used in Eq. (11).

To illustrate the use of the algorithm described in this section, we have synthetically generated a set of samples (corresponding to a common Gaussian distribution) with two clear outliers. This is shown in Fig. 5. The method defined above, is able to automatically detect the two outliers of the data as shown in the figure. These have been circled in Fig. 5 for clear view.

## 5. Experimental results

### 5.1. Face database and localization

To test the approach proposed in this contribution, we have collected a database of still images (which we used for training) and video sequences (which we used for testing). The training set consists of three neutral face images per person. An example of the training set for one of the subjects is shown in Fig. 6. The testing set includes two distinct types of sequences, each of which contains 40 frames. The first set of video sequences corresponds to nearly frontal view faces with random talking; Fig. 7. The second type of sequences corresponds to faces with orientations raging from (roughly) $-50°$ to $+50°$; Fig. 8. Our current database consists of fifty people.

Fig. 6. The three training images for one of the subjects in our database.

Fig. 7. A few frames of one of the video sequences with random talking.

Fig. 8. Some frames of a video sequence with faces at varying pose.

For our experimental results reported below, we have used the algorithm of Heisel et al. [18] to automatically localize the position of the eyes, mouth and nose in each frame. Once these facial features have been localized, the original image is rotated until obtaining a frontal view face where the line connecting the centers of the two eyes is parallel to the horizontal axis. The top and bottom of the face are assigned as a function of these detected features. Finally, the face is warped to a final standard face of 120 by 170 pixels. Fig. 9 shows the warping results for the images shown in Fig. 7.

The algorithm described in [18] was found to have an average localization error of about 16 by 16 pixels. In our experiments we used three Gaussians to represent each of the distributions of each class $G=3$, a subspace of 49 dimensions to represent each of these Gaussian distributions $e=49$, 5 dimensions to construct the view-based eigenspace (where we do estimation of pose) $r=5$, and two closest eigenspaces are then used to determine the pose of any given test image $d=2$. In our experiments, the optical flow was computed using the method of Black and Anandan [6].

### 5.2. Expression changes

The first set of video sequences described above (i.e. those with random talking), are used to test the robustness of our algorithm over expression changes and imprecise localized faces. For comparison purposes, we also show the results obtained with those methods that do recognition from stills. The first group of algorithms we will test, is the same as that defined in this paper but with the difference that now only one image will be used in the testing stage. We refer to these algorithms as L-$a$, where 'L' stands for local approach and (as before) $a=\{\text{PCA, ICA, LDA}\}$. The second group of algorithms will be the classical implementations of PCA, ICA and LDA (as defined in [39,1,3]). Their results will be labelled G-$a$, where now 'G' means global.

Note that while the L-$a$ methods should be robust to occlusions, expression and pose changes and errors of localization, the G-$a$ methods need not be. This is important because the difference in classification error we observe between the L-$a$ algorithms and those that use a video sequence will represent the gain obtained when using multiple images for testing.

We also note that since the test sequences contain more than one image, the L-$a$ and G-$a$ methods will produce different results depending on the actual frame used from them. To solve this problem, we will calculate the average recognition rate. This means, we will first obtain the recognition rate for each of the frames separately and, then, compute the average.

It is important to note that the approach defined in this paper could have also been formulated with a voting framework. In general, the use of probabilities is much preferred for the following reasons: (*A*) It is easier to detect outliers when one uses a probabilistic approach because these are commonly associated to large deviations of the parameters of the distribution; (*B*) We can use a confidence value to determine when a class has been correctly identified, whereas in voting we need to define a threshold (which is generally problem specific); (*C*) The probabilistic framework also allows us to select those images that are best for training and/or testing, while in the voting approach one needs to use them all.

However, in some cases, a simple voting approach (which uses the formulation defined in this paper) will suffice. This is the case in some, but not all, the test we have conducted and that are reported below. For this reason, we will also compare our algorithm to one which uses all the key components defined above but combines the classification result obtained in each of the image of the video sequence using a simple voting approach (in lieu of our probabilistic framework). This method

Fig. 9. Warped faces automatically obtained from the video sequence shown in Fig. 7.

should be robust to errors of localization, expression changes, occlusions and pose, because these are addressed by our formulation. Unfortunately, as we will latter see, the flexibility of the voting approach is not always sufficient to solve all these problems successfully.

The successful classification rates are shown in Fig. 10(a)–(c). In this figure, we show how the proposed algorithm behaves as a function of rank and cumulative match score and how they compare to the results given by the L-$a$ and G-$a$ algorithms. Rank means that the correct solution is within the R nearest neighbors and cumulative match score refers to the percentage of successfully recognized images [32].

The results in (a) are obtained with $a = PCA$, those in (b) with $a = ICA$ and those in (c) with $a = LDA$. The rank-1 match scores are shown in Fig. 10. As expected our algorithm outperforms those that only use a single image for testing. In this case, both, the probabilistic and voting, approaches perform equivalently.

### 5.3. Occlusions

We now want to study the amount of occlusion that the proposed algorithm can handle. To test this we defined the following procedure. The three neutral expression images were (again) used for training. Testing was done as above—using those video sequences with random talking. However, now, black squares of $s \times s$ pixels were randomly located in each of the images of these test sequences. Here, the recognition rate will not only depend on the size of the square, $s$, but also on the location of the square within the image. For this reason, we first varied the size of the square from a low of 10 to a maximum of 80 (which we vary at increments of $10 \times 10$ pixels) and randomly located the square in each frame of the video sequences 100 times (which gives us a total of 40,000 sequences). The results are summarized in Fig. 11(a–c). In this figure, the horizontal axis represents the size of the occluding square ($s \times s$) and the vertical axis the successful recognition rate. The average and standard deviation of the results obtained with PCA are shown in (a), those of ICA in (b) and those of LDA in (c). In each of these cases, we show the results obtained with the probabilistic framework and the voting approach defined above.

As expected the results of the voting approach are similar to those of the probabilistic algorithm for small occlusions. However, as the size of the occlusion is made larger, the difference in performance between the voting and probabilistic

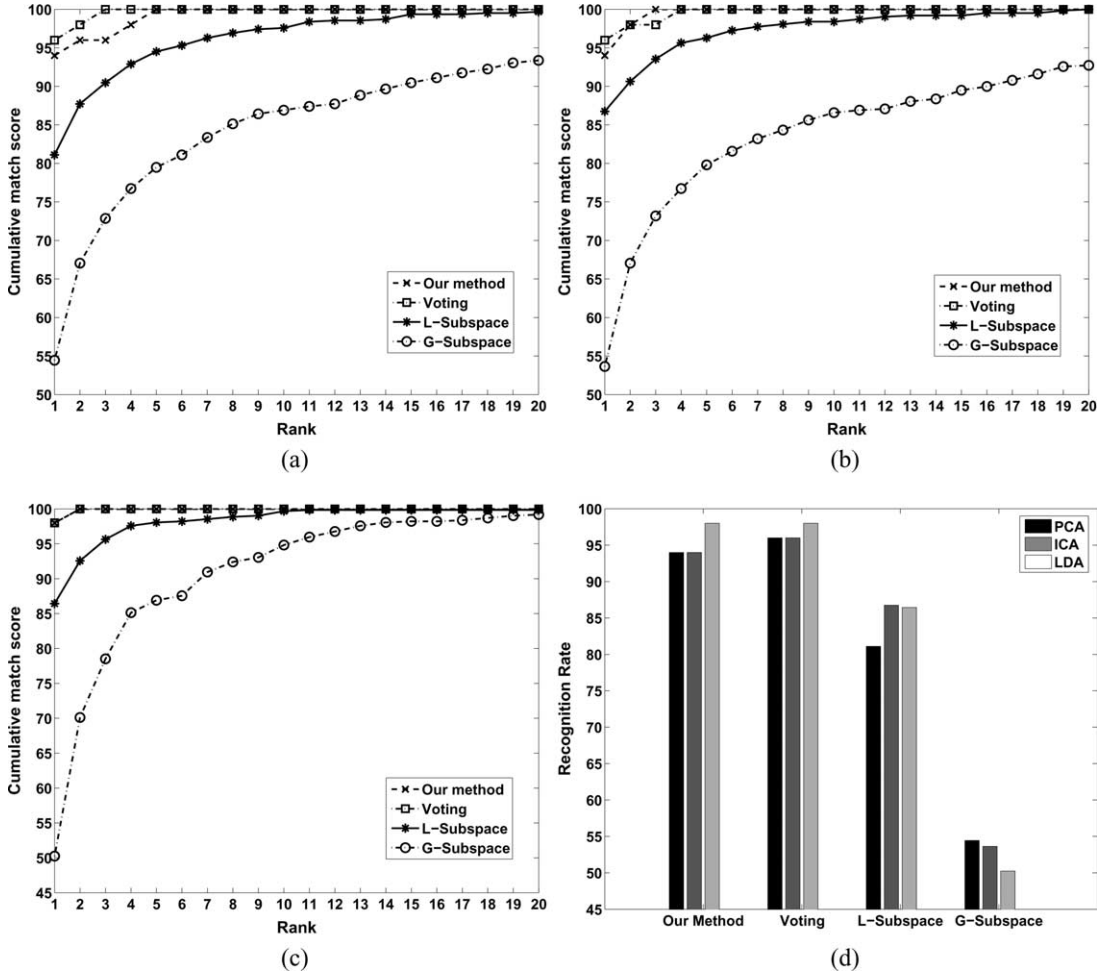


Fig. 10. Shown here are the results obtained with our probabilistic algorithm, our voting algorithm and two implementations of subspace techniques—one local, one global. (a)–(c) Show cumulative matching score using PCA, ICA and LDA, respectively. (e) Shows the rank-1 recognition rate for each of the methods.

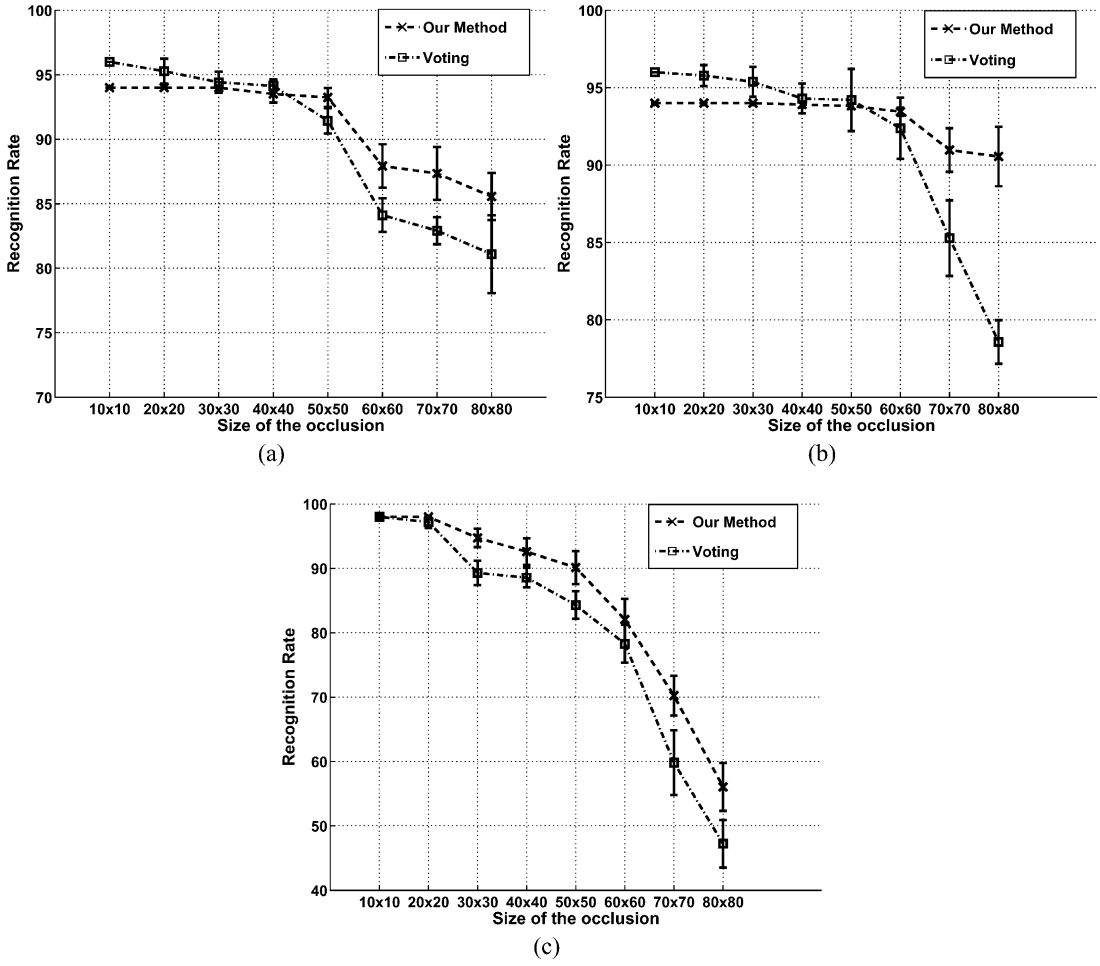


Fig. 11. Shown here is the reduction in recognition rate as a function of the size of the occluding square ($s \times s$). Results obtained with (a) PCA, (b) ICA and (c) LDA.

approaches increases. This is especially true for the cases where we used ICA, which is (at least in our experiments) the best representation.

Another interesting conclusion can be drawn from this experiment. While LDA produces slightly superior results when the faces are barely occluded, these results degrade rapidly when $s$ is larger than 30. Although interesting, this result should not be surprising. The problem with the subspaces generated by LDA is that they are too tuned to the training data. More specifically, LDA will select those features that best discriminate among classes. And, when these features get occluded our LDA representation will have less useful information to compensate for this lost. In contrast, PCA and ICA will find those features that are best to represent (describe) the 'appearance' of the face. When one area of the face is occluded, we can still use another. We see that PCA and ICA give very good results even when the occlusion is large. ICA being slightly superior for large occlusions; e.g. $s \geq 60$.

### 5.4. Variations in pose

In the two sections above, we have tested the robustness of our algorithm with regard to expression changes, occlusions and errors of localization. We now show how our algorithm performs on the second set of sequences—those that include pose changes. In this case, we will also compare our results to those obtained with the L-$a$ and G-$a$ methods. Note that while all methods use the same three (neutral expression) images for training, testing is done using either the whole sequence (for our method) or a single still (for the L- and G-$a$ algorithms).

In this experiment, the detection of the faces and corresponding facial features is done manually, because current algorithms cannot reliably detect all facial features from images with a rotation higher than 30 degrees.

The results of our test are summarized in Fig. 12(a)–(d). This figure also shows the results in a rank versus cumulative match score plot. The results obtained with $a$ = PCA are in (a), those of ICA in (b) and those of LDA in (c). The rank-1 match scores (for each of the methods) are in Fig. 12(d). Here, our probabilistic approach outperforms all the other methods. Note that, in this test, the voting cannot classify the data correctly. The capacity of the probabilistic approach to select those images that are best for recognition (by assigning higher probabilities) is key to this success.

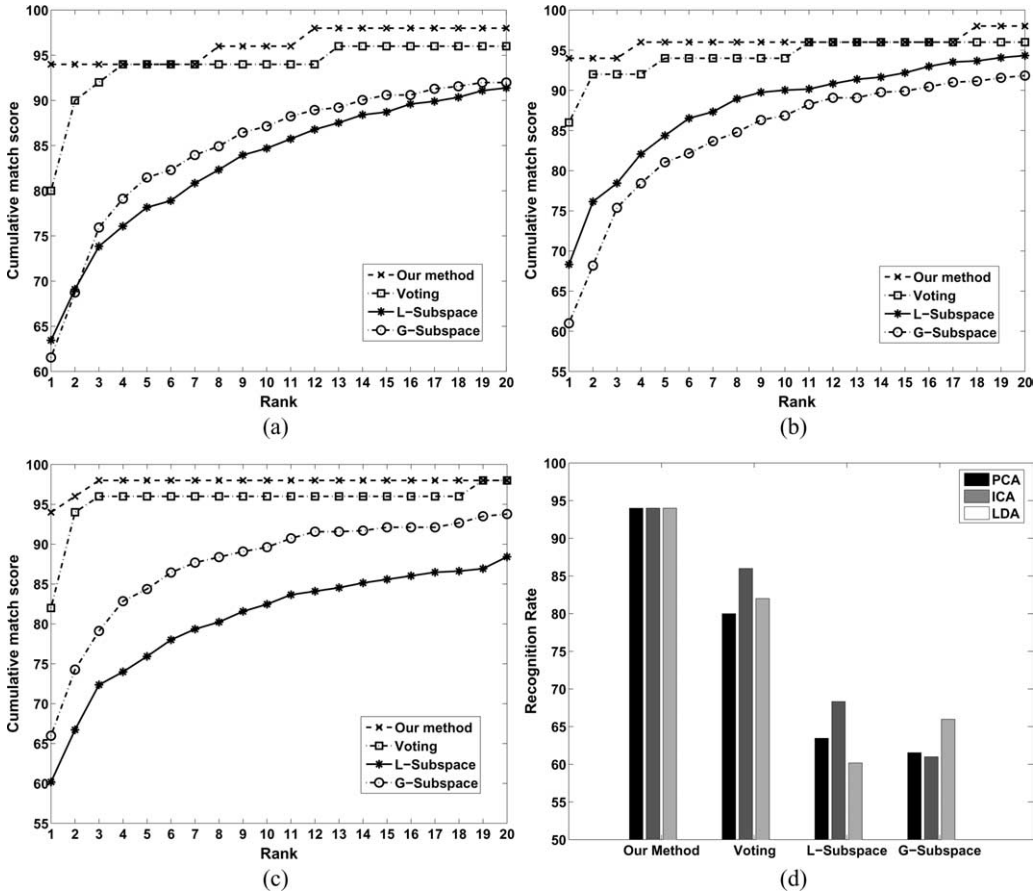


Fig. 12. Results obtained on the sequences with faces at different pose. (a)–(c) Show cumulative match score versus rank plots when using PCA, ICA and LDA as feature extractor. (e) Shows the rank-1 recognition rates for each of the methods.

Another interesting finding is that the results obtained with our probabilistic method are independent of the feature extraction algorithm used—in the sense that they are all equivalent. This is, however, not true for the other algorithms.

### 5.5. Discussion

The results reported above show that our probabilistic algorithm is more robust to the image variations summarized in

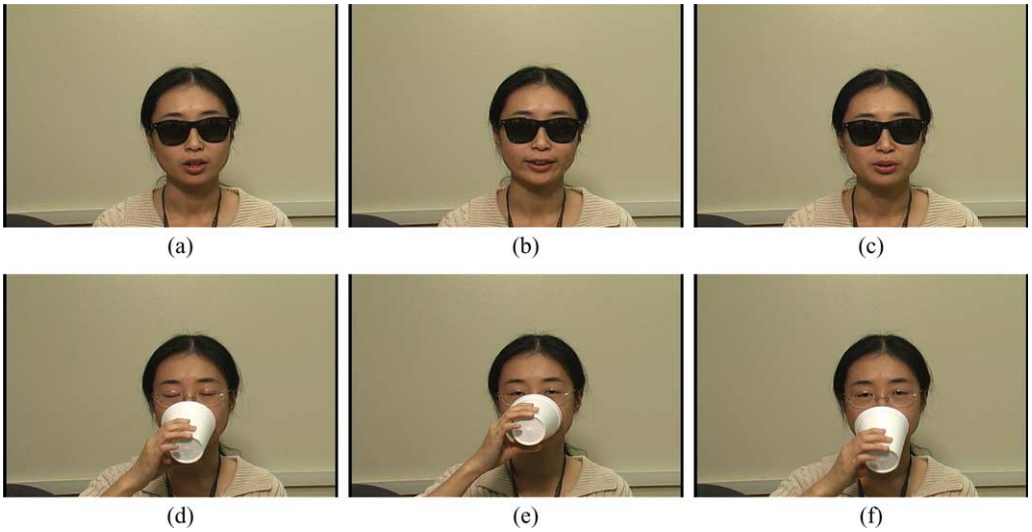


Fig. 13. Example images of the sequences with occlusions; sunglasses in (a)–(c) and drinking water in (d)–(f). These sequences were recorder 10 and 11 month after the ones used for training.

the introduction than those algorithms that use a single still for recognition and to those that use a voting approach. This is especially true, for large occlusions and pose variations.

As a final note, we would like to bring a word of caution to the positive comments presented above. While the method proposed in this paper has shown promise in solving the four main problems outlined in the introduction—errors of localization, occlusions, and expression and pose changes—from video sequences, much is yet to be accomplished before we can use such systems in unconstrained environments. The recognition of duplicate images (i.e. those images of the same people taken months or years apart) is a well-known example of this.

To test the potential use of our algorithm with duplicate images, we collected four additional video sequences form eleven of the fifty subjects in our database. These sequences were taken ten to eleven months after those used for training. The first two sequences are similar to the ones described above: (1) frontal faces with random talking, and (2) images at distinct poses. The other two correspond to subjects: (3) wearing large sunglasses and with random talking, and (4) drinking from a cup. While in (3) the top part of the face is occluded in each of the frames of the video sequence, in (4) the bottom part is the one that is not visible. A few selected images of the sequences in (3) and (4) for one of the subjects are shown in Fig. 13.

The successful recognition rates of our method (with $a =$ PCA) were: 100% for those sequences in (1) and (2), and around 82% for those in (3) and (4). Results with a much larger number of people and sequences will be necessary to reach any conclusive result. Such a database is not currently available. Some groups are currently working toward solving this problem.

## 6. Conclusions

This paper describes a weighted probabilistic approach that learns from a set of still images and then does recognition from video sequences. A main goal of this paper was to demonstrate that when one uses video sequences for testing, the results are significantly improved as compared to those obtained with one single still. Particularly, we have shown two main points. First, that the problems of occlusions, expression and pose changes, and errors of localization can be more easily solved when one uses a video sequence than when using a single still. And, second, that regardless of the method used to do feature extraction, the results always improve when we use multiple images. Another important conclusion is that while PCA and ICA are good representations to solve the problem of partial occlusions, LDA is only expected to work well when such occlusions are small. By closely examining our results, one can also conclude that while we may seem to correctly address the problems posed by occlusions, expressions and errors of localization, our test only explored the use of frontal face images. Further work should thus focus on the analysis of those features that are most invariant to pose changes. It would be interesting to know whether such features can be extracted more easily from video sequences or from stills. The collection of a database with a much larger set of people will also play an important role in future research.

## Acknowledgements

We would like to thank the reviewers for their comments. Thanks also go to the people of the CBCL lab at MIT for sending us the code of their face localization algorithm used in our experiments. This research was supported in part by the National Institutes of Health under grant R01-DC-005241.

## References

[1] M.S. Bartlett, Face Image Analysis by Unsupervised Learning, Kluwer, Dordecht, 2001.
[2] M.S. Bartlett, J.C. Hager, P. Ekman, T.J. Sejnowski, Measuring spatial expressions by computer image analysis, Psychophysiology 36 (1999) 253–263.
[3] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. Fisherfaces: recogniton using class specific linear projection, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 711–720.
[4] A.J. Bell, T.J. Sejnowski, An information-maximization approach to blind seperation and blind deconvolution, Neural Computation 7 (1995) 1129–1159.
[5] D. Beymer, T. Poggio, Face recognition from one example view, Science 272 (5250) (1996).
[6] M.J. Black, P. Anandan, The robust estimation of multiple motions: parametric and piece-wise smooth flow fields, Computer Vision and Image Understanding 63 (1) (1996) 75–104.
[7] T.F. Cootes, G.J. Edwards, C.J. Taylor, Active appearance models, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (6) (2001) 681–685.
[8] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likehood from incomplete data via the EM algorithm, Journal of Royal Statistical Society 30 (1) (1977) 1–38.
[9] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, second ed., Addison–Wesley, Reading, MA, 2002.
[10] G.J. Edwards, C.J. Taylor, T.F. Cootes, Improving identification performance by integrating evidence from sequence, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. I, 1999, pp. 486–491.
[11] K. Etemad, R. Chellapa, Discriminant analysis for recognition of human face images, Journal of Optical Society of American A 14 (8) (1997) 1724–1733.
[12] R.A. Fisher, The statistical utilization of multiple measurements, Annals of Eugenics 8 (1938) 376–386.
[13] K. Fukunaga, Introduction to Statistical Pattern Recognition, second ed., Academic Press, London, 1990.
[14] U. Gather, T. Hilker, C. Becker, A robustified version of sliced inverse regression, in: Proceedings of the Workshop on Statistical Methodology for the Sciences: Environmetrics and Genetics, 2001, pp. 147–157.
[15] S. Gong, A. Psarrou, I. Katsoulis, P. Palavouzis, Tracking and recognition of face sequence, in: European Workshop on Combined Real and Synthetic Image Processing for Broadcast and Video Production, 1994, pp. 96–112
[16] D. Gorodnichy, Recognizing faces in video requires approaches different from those developed for face recognition in photographs, in: Workshop on Enhancing Information Systems Security through Biometrics, Ottawa (Canada), 2004.
[17] P.L. Hallinan, G.G. Gordon, A.L. Yuille, P. Giblin, D. Mumford, Two- and Three-Dimensional Patterns of the Face, A.K. Peters, 1999.
[18] B. Heisel, T. Sere, M. Pontil, T. Vetter, T. Poggio, Categorization by learning and combining object parts, in: Proceedings of the NIPS, 2001.

[19] A.J. Howell, H. Buxton, Towards unconstrained face recognition from image sequences, in: Proceedings of the IEEE Conference on Automatic Face and Gesture Recognition, 1996, pp. 224–229.
[20] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, Wiley, 2001.
[21] G. Li, Z. Chen, Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and monte carlo, Journal of the American Statistical Association 80 (1985) 759–766.
[22] I.T. Jolliffe, Principal Component Analysis, second ed., Springer, 2002.
[23] K. Lee, J. Ho, M. Yang, D. Kriegman, Video-based face recognition using probabilistic appearance manifolds, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. I, 2003, pp. 313–320.
[24] H. Li, R. Chellappa, Face verification through tracking facial feature, Journal of the Optical Society of America A 8 (12) (2001).
[25] A.M. Martinez, Face image retrieval using HMMs, in: Proceedings of the IEEE Workshop on Content-Based Access of Images and Video Libraries, 1999, pp. 35–39.
[26] A.M. Martinez, Recognizing imprecisely localized, partially occuded and expression variant faces from a single sample per class, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (6) (2002) 748–763.
[27] A.M. Martinez, Matching expression variant faces, Vision Research 43 (9) (2003) 1047–1060.
[28] A.M. Martinez, Y. Zhang, Subset modeling of face localization error occlusion, and expression, in: R. Chellappa, W. Zhao (Eds.), Face Processing: Advanced Modeling and Methods, Academic Press, 2005.
[29] H. Murase, S.K. Nayar, Visual learning and recognition of 3-D object from appearance, International Journal of Computer Vision 14 (1995) 5–24.
[30] A. Pentland, B. Moghaddam, T. Starner, View-based and modular eigenspaces for face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1994.
[31] P.J. Phillips, H. Moon, P. Rauss, S.A. Rizvi, The FERET evaluation methodology for face-recognition algorithms, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (10) (2000) 1090–1104.
[32] Phillips P.J., Grother P., Micheals R.J., Blackburn D.M., Tabassi E., Bone M., Face Recognition Vendor Test 2002: Evaluation Report, Technical Report on NISTIR 6965, National Institute of Standards and Technology, 2003, Available at http://www.frvt.org.
[33] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, W. Worek, Overview of the face recognition grand challenge, in: Proceedings of the IEEE Compuetr Vision and Pattern Recognition, 2005.
[34] G. Shakhnarovich, J.W. Fisher, T. Darell, Face recognition from long-term observations, in: Proceedings of the European Conference on Computer Vision, 2002, pp. 851–868.
[35] L. Sirovich, M. Kirby, Low-dimensional procedure for the characterization of human faces, Journal of the Optical Society of America A 4 (1987) 519–524.
[36] S. Srinivasan, K.L. Boyer, Head pose estimation using view based eigenspaces, in: Proceedings of the International Conference on Pattern Recognition, 2002, pp. 302–305.
[37] K.-K. Sung, T. Poggio, Example-based learning for view-based human face detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (1) (1998) 39–51.
[38] M. Turk, A. Pentland, Eigenfaces for recogntion, Journal of Cognitive Neuroscience 3 (1) (1991) 71–86.
[39] T. Vetter, N.F. Troje, Separation of texture and shape in images of faces for image coding and synthesis, Journal of the Optical Society of America A 14 (9) (1997) 2152–2161.
[40] H. Wechsler, V. Kakkad, J. Huang, S. Gutta, V. Chen, Automatic video-based person authentication using the RBF network, in: Proceedings of the International Conference on Audio and Video-based Biometric Person Authentication, 1997, pp. 117–183.
[41] Y. Yacoob, L. Davis, Computing spatio-temporal representation of human faces, in: Proceedings of the Conference on Computer Vision and Pattern Recognition, 1994, pp. 70–75.
[42] W.Y. Zhao, R. Chellappa, A. Rosenfeld, J.P. Phillips, Face recognition: A literature survey, ACM Computing Surveys (2003).
[43] S. Zhou, V. Krueger, R. Chellappa, Probabilistic recognition of human faces from video, Computer Vision and Image Understanding 91 (2003) 214–245.