PERGAMON

Pattern Recognition 33 (2000) 1771–1782

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Bayesian face recognition

Baback Moghaddam[a,*], Tony Jebara[b], Alex Pentland[b]

[a] *Mitsubishi Electric Research Laboratory, 201 Broadway, 8th floor, Cambridge, MA 02139, USA*
[b] *Massachusettes Institute of Technology, Cambridge, MA 02139, USA*

Received 15 January 1999; received in revised form 28 July 1999; accepted 28 July 1999

Abstract

We propose a new technique for direct visual matching of images for the purposes of face recognition and image retrieval, using a *probabilistic* measure of similarity, based primarily on a Bayesian (MAP) analysis of image differences. The performance advantage of this probabilistic matching technique over standard Euclidean nearest-neighbor eigenface matching was demonstrated using results from DARPA's 1996 "FERET" face recognition competition, in which this Bayesian matching alogrithm was found to be the top performer. In addition, we derive a simple method of replacing costly computation of *nonlinear* (on-line) Bayesian similarity measures by inexpensive *linear* (off-line) subspace projections and simple Euclidean norms, thus resulting in a significant computational speed-up for implementation with very large databases. © 2000 Pattern Recognition Society. Published by Elsevier Science Ltd. All rights reserved.

Keywords: Face Recognition; Density estimation; Bayesian analysis; MAP/ML classification; Principal component analysis; Eigenfaces

1. Introduction

In computer vision, face recognition has a distinguished lineage going as far back as the 1960s with the work of Bledsoe [1]. This system (and many others like it) relied on the geometry of (manually extracted) fiducial points such as eye/nose/mouth corners and their spatial relationships (angles, length ratios, etc.). Kanade [2] was first to develop a fully automatic version of such a system. This "feature-based" paradigm persisted (or laid dormant) for nearly 30 years, with researchers often disappointed by the low recognition rates achieved even on small data sets. It was not until the 1980s that researchers began experimenting with visual representations, making use of the appearance or texture of facial images, often as raw 2D inputs to their systems. This new paradigm in face recognition gained further momentum due, in part, to the rapid advances in connectionist models in the 1980s which made possible face recognition systems such as the layered neural network systems of O'Toole et al. [3], Flemming and Cottrell [4] as well as the associative memory models used by Kohonen and Lehtio [5]. The debate on features vs. templates in face recognition was mostly settled by a comparative study by Brunelli and Poggio [6], in which template-based techniques proved significantly superior. In the 1990s, further developments in template or appearance-based techniques were prompted by the ground-breaking work of Kirby and Sirovich [7] with Karhunen–Loève Transform [8] of faces, which led to the principal component analysis (PCA) [9] "eigenface" technique of Turk and Pentland [10]. For a more comprehensive survey of face recognition techniques the reader is referred to Chellappa et al. [11].

The current state-of-the-art in face recognition is characterized (and to some extent dominated) by a family of subspace methods originated by Turk and Pentland's "eigenfaces" [10], which by now has become a de facto standard and a common performance benchmark in the field. Extensions of this technique include view-based and modular eigenspaces in Pentland et al. [12] and probabilistic subspace learning in Moghaddam and Pentland [13,14]. Examples of other subspace techniques include subspace mixtures by Frey and Huang [15], linear

[*] Corresponding author. Tel.: + 1-617-621-7524; fax: + 1-617-621-7500.
E-mail address: baback@merl.com (B. Moghaddam).

0031-3203/00/$20.00 © 2000 Pattern Recognition Society. Published by Elsevier Science Ltd. All rights reserved.
PII: S0031-3203(99)00179-X

discriminant analysis (LDA) as used by Etemad and Chellappa [16], the "Fisherface" technique of Belhumeur et al. [17], hierarchical discriminants used by Swets and Weng [18] and "evolutionary pursuit" of optimal subspaces by Liu and Wechsler [19] — all of which have proved equally (if not more) powerful than standard "eigenfaces".

Eigenspace techniques have also been applied to modeling the shape (as opposed to texture) of the face. Eigenspace coding of shape-normalized or "shape-free" faces, as suggested by Craw and Cameron [20], is now a standard pre-processing technique which can enhance performance when used in conjunction with shape information [21]. Lanitis et al. [22] have developed an automatic face-processing system with subspace models of both the shape and texture components, which can be used for recognition as well as expression, gender and pose classification. Additionally, subspace analysis has also been used for robust face detection [12,14,23], nonlinear facial interpolation [24], as well as visual learning for general object recognition [13,25,26].

2. A Bayesian approach

All of the face recognition systems cited above (indeed the majority of face recognition systems published in the open literature) rely on similarity metrics which are invariably based on Euclidean distance or normalized correlation, thus corresponding to standard "template-matching" — i.e., nearest-neighbor-based recognition. For example, in its simplest form, the similarity measure $S(I_1, I_2)$ between two facial images I_1 and I_2 can be set to be inversely proportional to the norm $\|I_1 - I_2\|$. Such a simple metric suffers from a major drawback: it does not exploit knowledge of which types of variation are critical (as opposed to incidental) in expressing similarity.

In this paper, we present a *probabilistic* similarity measure based on the Bayesian belief that the image intensity differences, denoted by $\Delta = I_1 - I_2$, are characteristic of typical variations in appearance of an individual. In particular, we define two classes of facial image variations: *intrapersonal* variations Ω_I (corresponding, for example, to different facial expressions of the *same* individual) and *extrapersonal* variations Ω_E (corresponding to variations between *different* individuals). Our similarity measure is then expressed in terms of the probability

$$S(I_1, I_2) = P(\Delta \in \Omega_I) = P(\Omega_I | \Delta), \tag{1}$$

where $P(\Omega_I|\Delta)$ is the a posteriori probability given by Bayes rule, using estimates of the likelihoods $P(\Delta|\Omega_I)$ and $P(\Delta|\Omega_E)$. These likelihoods are derived from training data using an efficient subspace method for density estimation of high-dimensional data [14], briefly reviewed in Section 3.1.

We believe that our Bayesian approach to face recognition is possibly the first instance of a non-Euclidean similarity measure used in face recognition [27–30]. Furthermore, our method can be viewed as a generalized nonlinear extension of linear discriminant analysis (LDA) [16,18] or "FisherFace" techniques [17] for face recognition. Moreover, the mechanics of Bayesian matching has computational and storage advantages over most linear methods for large databases. For example, as shown in Section 3.2, one need only store a single image of an individual in the database.

3. Probabilistic similarity measures

In previous work [27,31,32], we used Bayesian analysis of various types of facial appearance models to characterize the observed variations. Three different inter-image representations were analyzed using the binary formulation (Ω_I- and Ω_E-type variation): XYI-warp modal deformation spectra [27,31,32], XY-warp optical flow fields [27,31] and a simplified I-(intensity)-only image-based differences [27,29]. In this paper we focus on the latter representation only — the normalized intensity difference between two facial images which we refer to as the Δ vector.

We define two distinct and mutually exclusive classes: Ω_I representing *intrapersonal* variations between multiple images of the same individual (e.g., with different expressions and lighting conditions), and Ω_E representing *extrapersonal* variations in matching two different individuals. We will assume that both classes are Gaussian-distributed and seek to obtain estimates of the likelihood functions $P(\Delta|\Omega_I)$ and $P(\Delta|\Omega_E)$ for a given intensity difference $\Delta = I_1 - I_2$.

Given these likelihoods we can evaluate a similarity score $S(I_1, I_2)$ between a pair of images directly in terms of the intrapersonal a posteriori probability as given by Bayes rule:

$$S(I_1, I_2) = \frac{P(\Delta|\Omega_I)P(\Omega_I)}{P(\Delta|\Omega_I)P(\Omega_I) + P(\Delta|\Omega_E)P(\Omega_E)}, \tag{2}$$

where the priors $P(\Omega)$ can be set to reflect specific operating conditions (e.g., number of test images vs. the size of the database) or other sources of a priori knowledge regarding the two images being matched. Note that this particular Bayesian formulation casts the standard face recognition task (essentially an M-ary classification problem for M individuals) into a *binary* pattern classification problem with Ω_I and Ω_E. This simpler problem is then solved using the maximum a posteriori (MAP) rule — i.e., two images are determined to belong to the same individual if $P(\Omega_I|\Delta) > P(\Omega_E|\Delta)$, or equivalently, if $S(I_1, I_2) > \frac{1}{2}$.

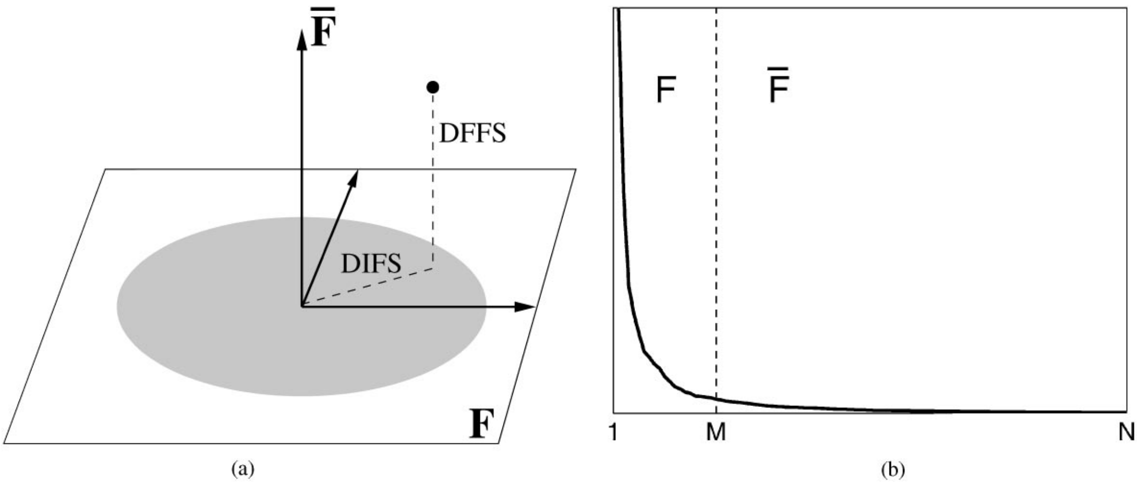

Fig. 1. (a) Decomposition of $\mathscr{R}^N$ into the principal subspace F and its orthogonal complement $\bar{F}$ for a Gaussian density, (b) a typical eigenvalue spectrum and its division into the two orthogonal subspaces.

An alternative probabilistic similarity measure can be defined in simpler form using the intrapersonal likelihood alone,

$$S' = P(\Delta|\Omega_I), \tag{3}$$

thus leading to *maximum likelihood* (ML) recognition as opposed to the MAP recognition in Eq. (2). Our experimental results in Section 4 indicate that this simplified ML measure can be almost as effective as its MAP counterpart in most cases.

3.1. Subspace density estimation

One difficulty with this approach is that the intensity difference vector is very high-dimensional, with $\Delta \in \mathscr{R}^N$ with N typically of $O(10^4)$. Therefore we almost always lack sufficient independent training samples to compute reliable second-order statistics for the likelihood densities (i.e., singular covariance matrices will result). Even if we were able to estimate these statistics, the computational cost of evaluating the likelihoods is formidable. Furthermore, this computation would be highly inefficient since the *intrinsic* dimensionality or major degrees-of-freedom of Δ is likely to be significantly smaller than N.

To deal with the high dimensionality of Δ, we make use of the efficient density estimation method proposed by Moghaddam and Pentland [13,14] which divides the vector space $\mathscr{R}^N$ into two complementary subspaces using an eigenspace decomposition. This method relies on a principal components analysis (PCA) [9] to form a low-dimensional estimate of the complete likelihood which can be evaluated using only the first M principal components, where $M \ll N$.

This decomposition is illustrated in Fig. 1 which shows an orthogonal decomposition of the vector space $\mathscr{R}^N$ into two mutually exclusive subspaces: the principal subspace F containing the first M principal components and its orthogonal complement F, which contains the residual of the expansion. The component in the orthogonal subspace F is the so-called "distance-from-feature-space" (DFFS), a Euclidean distance equivalent to the PCA residual error. The component of Δ which lies *in* the feature space F is referred to as the "distance-in-feature-space" (DIFS) and is a *Mahalanobis* distance for Gaussian densities.

As shown in Refs. [13,14], the complete likelihood estimate can be written as the product of two independent marginal Gaussian densities

$$\hat{P}(\Delta|\Omega) = \left[\exp\frac{\left(-\frac{1}{2}\sum_{i=1}^{M} y_i^2/\lambda_i\right)}{(2\pi)^{M/2}\prod_{i=1}^{M}\lambda_i^{1/2}}\right]\left[\frac{\exp\left(-\varepsilon^2(\Delta)/2\rho\right)}{(2\pi\rho)^{(N-M)/2}}\right]$$

$$= P_F(\Delta|\Omega)\hat{P}_F(\Delta|\Omega), \tag{4}$$

where $P_F(\Delta|\Omega)$ is the true marginal density in F, $\hat{P}_{\bar{F}}(\Delta|\Omega)$ is the estimated marginal density in the orthogonal complement $\bar{F}$, y_i are the principal components and $\varepsilon^2(\Delta)$ is the residual (DFFS). The optimal value for the weighting parameter ρ — found by minimizing cross-entropy — is simply the average of the F eigenvalues[1]

$$\rho = \frac{1}{N-M}\sum_{i=M+1}^{N}\lambda_i. \tag{5}$$

[1] Tipping and Bishop [33] have since derived the same estimator for ρ by showing that it's a saddle point of the likelihood for a latent variable model.

We note that in actual practice, the majority of the $\bar{F}$ eigenvalues are unknown but *can* be estimated, for example, by fitting a nonlinear function to the available portion of the eigenvalue spectrum and estimating the average of the eigenvalues beyond the principal subspace.

3.2. Efficient similarity computation

Consider a feature space of Δ vectors, the differences between two images (I_j and I_k). The two classes of interest in this space correspond to intrapersonal and extrapersonal variations and each is modeled as a high-dimensional Gaussian density

$$P(\Delta|\Omega_E) = \frac{\mathrm{e}^{-\frac{1}{2}\Delta^{\mathrm{T}}\sum_E^{-1}\Delta}}{(2\pi)^{D/2}|\sum_E|^{1/2}},$$

$$P(\Delta|\Omega_I) = \frac{\mathrm{e}^{-\frac{1}{2}\Delta^{\mathrm{T}}\sum_I^{-1}\Delta}}{(2\pi)^{D/2}|\sum_I|^{1/2}}. \tag{6}$$

The densities are zero-mean since for each $\Delta = I_j - I_k$ there exists a $\Delta = I_k - I_j$. Since these distributions are known to occupy a principal subspace of image space (face-space), only the principal eigenvectors of the Gaussian densities are relevant for modeling. These densities are used to evaluate the *similarity score* in Eq. (2) in accordance with the density estimate in Eq. (4).

Computing the similarity score involves first subtracting a candidate image I_j from a database entry I_k. The resulting Δ is then projected onto the principal eigenvectors of both extrapersonal and intrapersonal Gaussians. The exponentials are then evaluated, normalized and combined as likelihoods in Eq. (2). This operation is iterated over all members of the database (many I_k images) until the maximum score is found (i.e., the match). Thus, for large databases, this evaluation is rather expensive.

However, these compuations can be greatly simplified by offline transformations. To compute the likelihoods $P(\Delta|\Omega_I)$ and $P(\Delta|\Omega_E)$ we pre-process the I_k images with *whitening* transformations and consequently every image is stored as two vectors of whitened subspace coefficients; $\boldsymbol{i}$ for intrapersonal and $\boldsymbol{e}$ for extrapersonal

$$\boldsymbol{i}_j = \Lambda_I^{-1/2} V_I I_j \quad \boldsymbol{e}_j = \Lambda_E^{-1/2} V_E I_j \tag{7}$$

where, Λ and V are matrices of the largest eigenvalues and eigenvectors of $\sum_E$ or $\sum_I$, with subspace dimensionalities of M_I and M_E, respectively. After this pre-processing and with the normalizing denominators pre-computed, evaluating the likelihoods is reduced to computing simple Euclidean distances for the exponents

$$P(\Delta|\Omega_E) = \frac{\mathrm{e}^{-1/2\|\boldsymbol{e}_j - \boldsymbol{e}_k\|^2}}{(2\pi)^{D/2}|\sum_E|^{1/2}},$$

$$P(\Delta|\Omega_I) = \frac{\boldsymbol{e}^{-1/2\|\boldsymbol{i}_j - \boldsymbol{i}_k\|^2}}{(2\pi)^{D/2}|\sum_I|^{1/2}}. \tag{8}$$

These likelihoods are then used to compute the MAP similarity S in Eq. (2). Since the Euclidean distances in the exponents of Eq. (8) are of dimensions M_I and M_E for the $\boldsymbol{i}$ and $\boldsymbol{e}$ vectors, respectively, only $2 \times (M_I + M_E)$ arithmetic operations are required for each similarity computation. In this manner, one avoids unnecessary and repeated image differencing and online projections. The ML similarity matching based on Eq. (3) is even simpler to implement in this framework, since only the intra-personal class is evaluated, leading to the simplified similarity measure, computed using just the $\boldsymbol{i}$ vectors alone

$$S' = P(\Delta|\Omega_I) = \frac{\mathrm{e}^{-1/2\|\boldsymbol{i}_j - \boldsymbol{i}_k\|^2}}{(2\pi)^{D/2}|\sum_I|^{1/2}}. \tag{9}$$

4. Experiments

To test our recognition strategy we used a collection of images from the ARPA FERET face database. This collection of images consists of hard recognition cases that have proven difficult for most face recognition algorithms previously tested on the FERET database. The difficulty posed by this data set appears to stem from the fact that the images were taken at different times, at different locations, and under different imaging conditions. The set of images consists of pairs of frontal-views (FA/FB) which are divided into two subsets: the "gallery" (training set) and the "probes" (testing set). The gallery images consisted of 74 pairs of images (2 per individual) and the probe set consisted of 38 pairs of images, corresponding to a subset of the gallery members. The probe and gallery datasets were captured a week apart and exhibit differences in clothing, hair and lighting (see Fig. 2).

The front end to our system consists of an automatic face-processing module which extracts faces from the input image and normalizes for translation, scale as well as slight rotations (both in-plane and out-of-plane). This system is described in detail in Refs. [13,14] and uses maximum-likelihood estimation of object location (in this case the position and scale of a face and the location of individual facial features) to geometrically align faces into standard normalized form as shown in Fig. 3. All the faces in our experiments were geometrically aligned and normalized in this manner prior to further analysis.

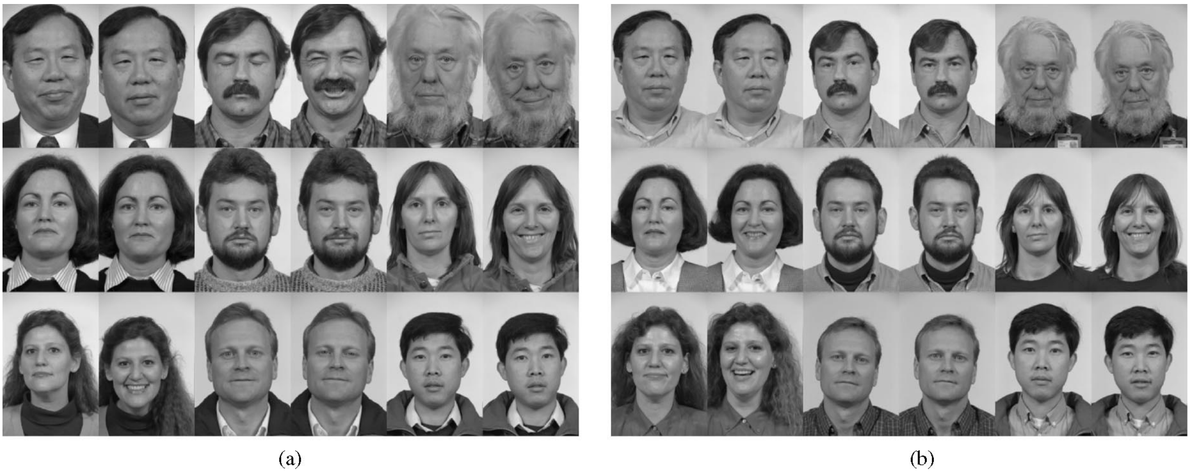

Fig. 2. Examples of FERET frontal-view image pairs used for (a) the Gallery set (training) and (b) the Probe set (testing).

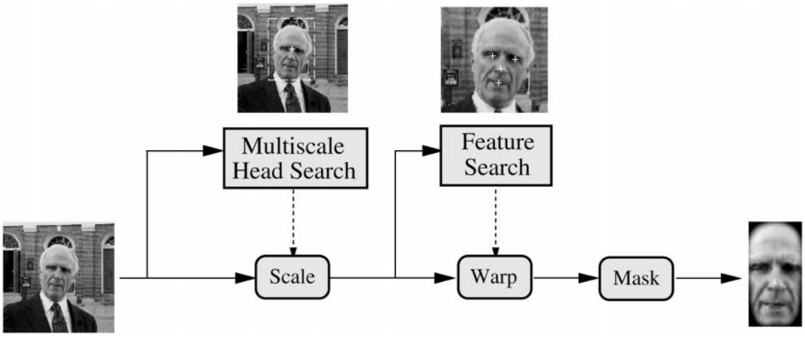

Fig. 3. Face alignment system.

4.1. Eigenface matching

As a baseline comparison, we first used an eigenface matching technique for recognition [10]. The normalized images from the gallery and the probe set were projected onto eigenfaces similar to those shown in Fig. 4. A nearest-neighbor rule based on a Euclidean distance was then used to match each probe image to a gallery image. We note that this corresponds to a generalized template-matching method which uses a Euclidean norm restricted to the principal subspace of the data. We should also add that these eigenfaces represent the principal components of an entirely different set of images — i.e., none of the individuals in the gallery or probe sets were used in obtaining these eigenvectors. In other words, neither the gallery nor the probe sets were part of the "training set". The rank-1 recognition rate obtained with this method was found to be 84% (64 correct matches out of 76), and the correct match was always in the top 10 nearest neighbors.

4.2. Bayesian matching

For our probabilistic algorithm, we first gathered training data by computing the intensity differences for a training subset of 74 intrapersonal differences (by matching the two views of every individual in the gallery) and a random subset of 296 extrapersonal differences (by matching images of *different* individuals in the gallery), corresponding to the classes Ω_I and Ω_E, respectively, and performed a separate PCA analysis on each.

It is interesting to consider how these two classes are distributed, for example, are they linearly separable or embedded distributions? One simple method of visualizing this is to plot their mutual principal components — i.e., perform PCA on the *combined* dataset and project

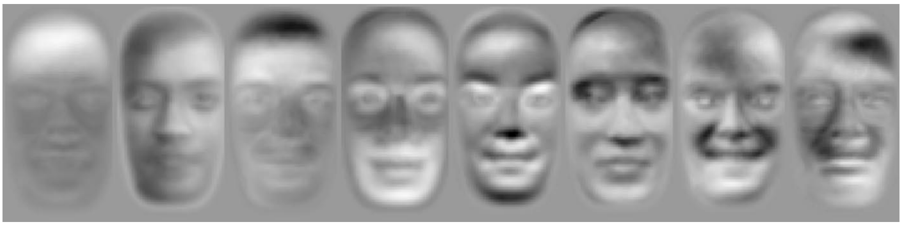

Fig. 4. Standard eigenfaces.

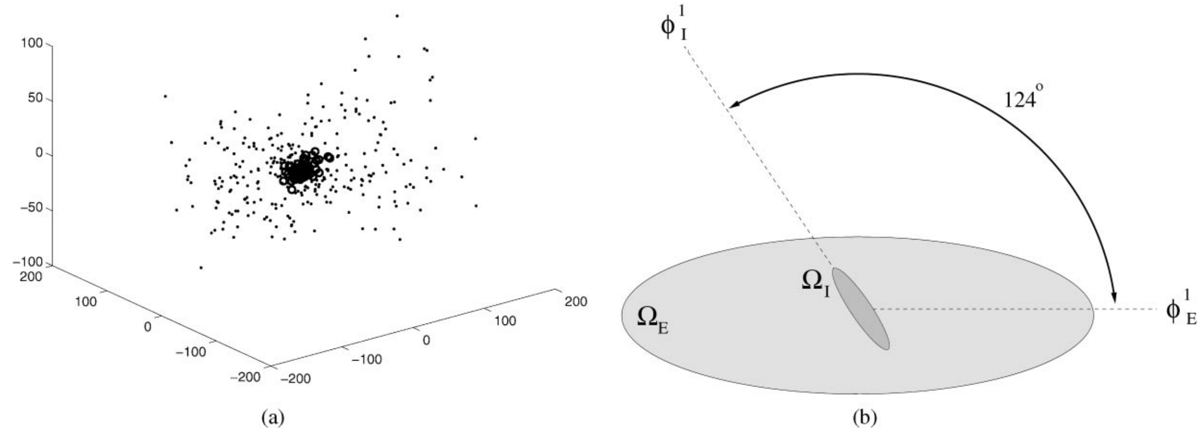

Fig. 5. (a) Distribution of the two classes in the first three principal components (circles for Ω_I, dots for Ω_E) and (b) schematic representation of the two distributions showing orientation difference between the corresponding principal eigenvectors.

each vector onto the principal eigenvectors. Such a visualization is shown in Fig. 5(a) which is a 3D scatter plot of the first three principal components. This plot shows what appears to be two completely enmeshed distributions, both having near-zero means and differing primarily in the amount of scatter, with Ω_I displaying smaller intensity differences as expected. It therefore appears that one cannot reliably distinguish low-amplitude extrapersonal differences (of which there are many) from intrapersonal ones.

However, direct visual interpretation of Fig. 5(a) is misleading since we are essentially dealing with low-dimensional (or "flattened") hyper-ellipsoids which are intersecting near the origin of a very high-dimensional space. The key distinguishing factor between the two distributions is their relative orientation. We can easily determine this relative orientation by performing a separate PCA on each class and computing the dot product of their respective first eigenvectors. This analysis yields the cosine of the angle between the major axes of the two hyper-ellipsoids, which was found to be 124°, thus indicating that the relative orientations of the two hyper-ellipsoids are quite different. Fig. 5(b) is a schematic illustration of the geometry of this configuration, where the hyper-ellipsoids have been drawn to approximate scale using the corresponding eigenvalues.

4.3. Dual eigenfaces

We note that the two mutually exclusive classes Ω_I and Ω_E correspond to a "dual" set of eigenfaces as shown in Fig. 6. Note that the intrapersonal variations shown in Fig. 6(a) represent subtle variations due mostly to expression changes (and lighting) whereas the extrapersonal variations in Fig. 6(b) are more representative of standard variations such as hair color, facial hair and glasses. In fact, these extrapersonal eigenfaces are qualitatively similar to the standard eigenfaces shown in Fig. 4. This supports the basic intuition that intensity differences of the extrapersonal type span a larger vector space similar to the volume of facespace spanned by standard eigenfaces, whereas the *intrapersonal* eigenspace corresponds to a more tightly constrained subspace. It is the representation of this intrapersonal subspace that is the critical component of a probabilistic measure of facial similarity. In fact our experiments with a larger set of FERET

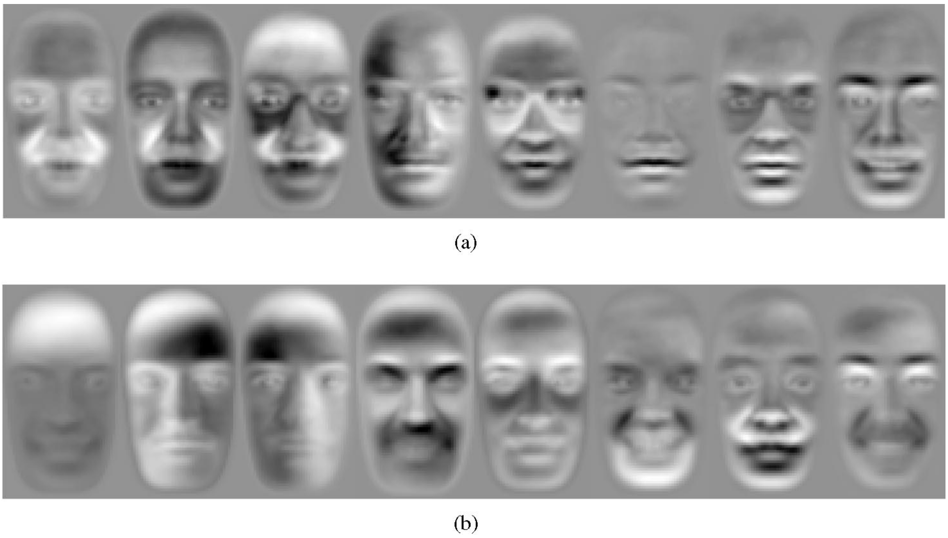

Fig. 6. “Dual” Eigenfaces: (a) Intrapersonal, (b) Extrapersonal.

images have shown that this intrapersonal eigenspace alone is sufficient for a simplified *maximum-likelihood* measure of similarity (see Section 4.4).

Note that since these classes are not linearly separable (they are both *zero-mean*), simple linear discriminant techniques (e.g., using hyperplanes) cannot be used with any degree of reliability. The proper decision surface is inherently nonlinear (hyperquadratic under the Gaussian assumption) and is best defined in terms of the a posteriori probabilities — i.e., by the equality $P(\Omega_I|\Delta) = P(\Omega_E|\Delta)$. Fortunately, the optimal discriminant surface is automatically implemented when invoking a MAP classification rule.

Having computed the two sets of training Δ's, we computed their likelihood estimates $P(\Delta|\Omega_I)$ and $P(\Delta|\Omega_E)$ using the susbspace method [13,14] described in Section 3.1. We used principal subspace dimensions of $M_I = 10$ and $M_E = 30$ for Ω_I and Ω_E, respectively. These density estimates were then used with a default setting of equal priors, $P(\Omega_I) = P(\Omega_E)$, to evaluate the a posteriori intrapersonal probability $P(\Omega_I|\Delta)$. This similarity was computed for each probe–gallery pair and used to rank the best matches accordingly. This probabilistic ranking yielded an improved rank-1 recognition rate of 89.5%. Furthermore, out of the 608 extrapersonal warps performed in this recognition experiment, only 2% (11) were misclassified as being intrapersonal — i.e., with $P(\Omega_I|\Delta) > P(\Omega_E|\Delta)$.

4.4. The 1996 FERET competition

This Bayesian approach to recognition has produced a significant improvement over the accuracy obtained with a standard eigenface nearest-neighbor matching rule. The probabilistic similarity measure was used in the September 1996 FERET competition (with subspace dimensionalities of $M_I = M_E = 125$) and was found to be the top-performing system by a typical margin of 10–20% over the other competing algorithms [34]. Fig. 7 shows the results of this test on a gallery of ≈ 1200 individuals. Note that rank-1 recognition rate is $\approx 95\%$ and significantly higher than the other competitors. In

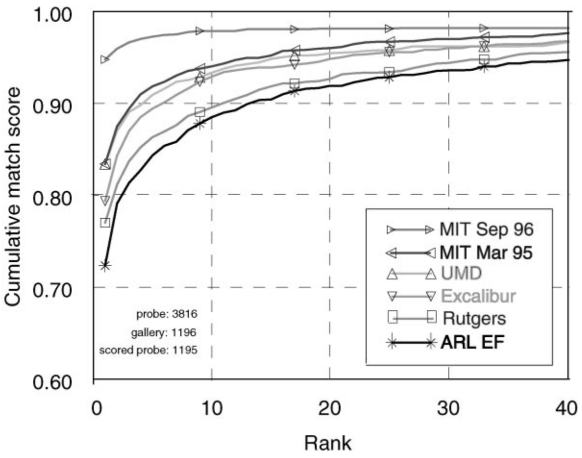

Fig. 7. Cumulative recognition rates for frontal FA/FB views for the competing algorithms in the FERET 1996 test. The top curve (labeled “MIT Sep 96”) corresponds to our Bayesian matching technique. Note that second placed is standard eigenface matching (labeled “MIT Mar 95”).

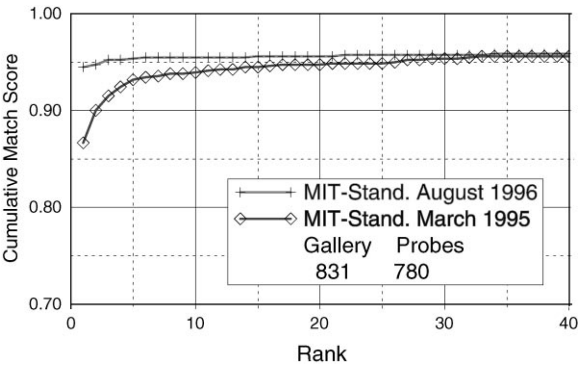

Fig. 8. Cumulative recognition rates with standard eigenface matching (bottom) and the newer Bayesian similarity metric (top).

fact, the next best system is our own implementation of standard eigenfaces. Fig. 8 highlights the performance difference between standard eigenfaces and the Bayesian method from a smaller test set of 800 + individuals. Note the 10% gain in performance afforded by the new Bayesian similarity measure which has effectively *halved* the error rate of eigenface matching.

As suggested in Section 3, a simplified similarity measure using only the *intrapersonal* eigenfaces can be used to obtain the *maximum-likelihood* (ML) similarity measure as defined in Eq. (3) and used instead of the *maximum* a posteriori (MAP) measure in Eq. (2). Although this simplified ML measure was not officially FERET tested, our experiments with a database of approximately 2000 faces have shown that using S' instead of S results in only a minor (2–3%) deficit in the recognition rate while cutting the computational cost by a factor of 2.

4.5. Eigenface vs. Bayesian matching

It is interesting to compare the computational protocol of standard Euclidean eigenfaces with the new probabilistic similarity. This is shown in Fig. 9 which illustrates the signal flow graphs for the two methods. With eigenface matching, both the probe and gallery images are pre-projected onto a single "universal" set of eigenfaces, after which their respective principal components are differenced and normed to compute a Euclidean distance metric as the basis of a similarity score. With probabilistic matching on the other hand, the probe and gallery images are **first** differenced and then projected onto **two** sets of eigenfaces which are used to compute the likelihoods $P(\Delta|\Omega_I)$ and $P(\Delta|\Omega_E)$, from which the a posteriori probability $P(\Omega_I|\Delta)$ is computed by application of Bayes rule as in Eq. (2). Alternatively, the likelihood $P(\Delta|\Omega_I)$ alone can be computed to form the simplified similarity in Eq. (3). As noted in the previous section, use of S' instead of S reduces the computational requirements by a factor of two, while only compromising the overall recognition rate by a few percentage points.[2]

Finally, we note that the computation of either MAP/ML similarity measures can be greatly simplified using the derivations shown in Section 3.2. This reformulation yields an exact remapping of the probabilistic similarity score without requiring repeated image-differencing and eigenspace projections. The most desirable aspect of this simplification is that the *nonlinear* matching of two images can be carried out in terms of simple Euclidean norms of their whitened feature vectors which are pre-computed off-line. As pointed out in Section 3.2, this is particularly appealing when working with large galleries of images and results in a significant online computational speedup.

5. Discussion

It remains an open research question as to whether the proposed Bayesian approach can be used to model larger variations in facial appearance other than expression or lighting. In particular, pose and facial "decorations" (e.g., glasses and beards) are complicating factors which are commonly encountered in realistic settings. With regards to the latter, we have been able to correctly recognize variations with/without regular glasses (not dark sunglasses) — in fact a sizeable percentage of the intrapersonal training set used in our experiments consisted of just such variations. Moderate recognition performance can also be achieved with simpler techniques like eigenfaces in the case of transparent eye-ware. Naturally, most face recognition systems can be fooled by sunglasses and significant variations in facial hair (beards).

This is not to say that one cannot — in principle — incorporate such gross variations as pose and extreme decorations into one comprehensive intrapersonal training set and hope to become invariant to them. But in our experience, this significantly increases the dimensionality of the subspaces and in essence *dilutes* the density models, rendering them ineffective. One preferred approach for dealing with large pose variations is the view-based *multiple model* method described in Ref. [14] whereby variable-pose recognition is delegated to multiple "experts" each of which is "tuned" to its own limited domain. For example, in earlier work [35], metric eigenface matching of a small set of variable-pose FERET images, consisting of $\{\text{frontal}, \pm \text{half}, \pm \text{profile}\}$ views yielded recognition

[2] Note that Fig. 9(b) shows the *conceptual* architecture of Bayesian matching and not the actual implementation used as detailed in Section 3.2.

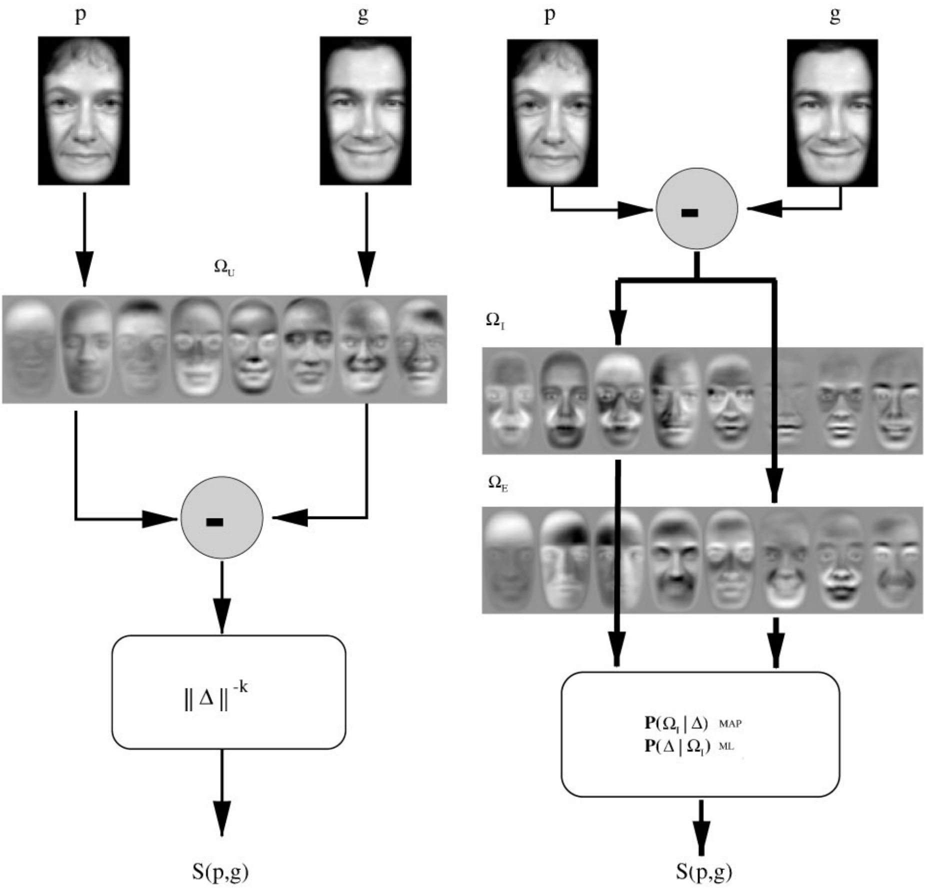

Fig. 9. Operational signal flow diagrams for (a) Eigenface similarity and (b) Probabilistic similarity.

rates of {99, 85, 69%}, respectively. However, it was found that cross-pose performance (train on one view, test on another) declined to $\approx 30\%$ with a mere $\approx 22\%$ change in pose. Therefore, the inability of a single view to generalize to other views indicates that multiple-model techniques [14] are a better way to tackle this problem.

6. Conclusions

The performance advantage of our probabilistic matching technique was demonstrated using both a small database (internally tested) as well as a large (1100 +) database with an independent double-blind test as part of ARPA's September 1996 "FERET" competition, in which Bayesian similarity out-performed all competing algorithms (at least one of which was using an LDA/Fisher type method). We believe that these results clearly demonstrate the superior performance of probabilistic matching over eigenface, LDA/Fisher and other existing Euclidean techniques.

This probabilistic framework is particularly advantageous in that the intra/extra density estimates explicitly characterize the type of appearance variations which are critical in formulating a meaningful measure of similarity. For example, the appearance variations corresponding to facial expression changes or lighting (which may have large image-difference norms) are, in fact, *irrelevant* when the measure of similarity is to be based on *identity*. The subspace density estimation method used for representing these classes thus corresponds to a *learning* method for discovering the principal modes of variation important to the recognition task. Consequently, only a single image of an individual can be used for recognition, thus reducing the storage cost with large databases.

Furthermore, by equating similarity with the a posteriori probability we obtain an optimal nonlinear decision rule for matching and recognition. This aspect of our approach differs significantly from methods which use linear discriminant analysis for recognition (e.g. [16,18]). This Bayesian method can be viewed as a generalized nonlinear (quadratic) version of linear discriminant analysis (LDA) [16] or "FisherFace" techniques [17]. The computational advantage of our approach is that there is no need to compute and store an eigenspace for each individual in the gallery (as required with pure LDA). One (or at most two) global eigenspaces are sufficient for probabilistic matching and therefore storage and computational costs are fixed and do not increase with the size of the training set (as is possible with LDA/Fisher methods).

The results obtained with the simplified ML similarity measure (S' in Eq. (3)) suggest a computationally equivalent yet superior alternative to standard eigenface matching. In other words, a likelihood similarity based on the intrapersonal density $P(\Delta|\Omega_I)$ alone is far superior to nearest-neighbor matching in eigenspace, while essentially requiring the same number of projections. However, for completeness (and slightly better performance) one should use the a posteriori similarity (S in Eq. (2)) at twice the computational cost of standard eigenfaces. Finally, in Section 3.2 we derived an efficient technique for computing (nonlinear) MAP/ML similarity scores using simple (linear) projections and Euclidean norms, making this method appealing in terms of computational simplicity and ease of implementation.

7. Summary

We have proposed a novel technique for direct visual matching of images for the purposes of recognition and search in a large face database. Specifically, we have argued in favor of a *probabilistic* measure of similarity, in contrast to simpler methods which are based on standard L_2 norms. The proposed similarity measure is based on a Bayesian analysis of image intensity differences: we model two mutually exclusive classes of variation between two face images: *intra-personal* (variations in appearance of the same individual, due to different expressions or lighting, for example) and *extra-personal* (variations in appearance due to different identity). The high-dimensional probability density functions for each respective class are then obtained from available training data using an efficient and optimal eigenspace density estimation technique and subsequently used to compute a similarity measure based on the a posteriori probability of membership in the *intra-personal* class. This posterior probability is then used to rank and find the best matches in the database. The performance advantage of our probabilistic matching technique has been demostrated using both a small database (internally tested) as well as large (1200 individual) database with an independent double-blind test as part of ARPA's September 1996 "FERET" competition, in which our Bayesian similarity matching technique out-performed all competing algorithms by at least 10% margin in recognition rate. This probabilistic framework is particularly advantageous in that the intra/extra density estimates explicity characterize the type of appearance variations which are critical in formulating a meaningful measure of similarity. For example, the intensity differences corresponding to facial expression changes (which may have high image-difference norms) are, in fact, *irrelevant* when the measure of similarity is to be based on *identity*. The subspace density estimation method used for representing these classes thus corresponds to a *learning* method for discovering the principal modes of variation important to the classification task.

References

[1] W.W. Bledsoe, The model method in facial recognition, Technical Report PRI:15, Panoramic Research Inc., Palo Alto, CA, August 1966.

[2] T. Kanade, Picture processing by computer complex and recognition of human faces, Technical Report, Kyoto University, Department of Information Science, 1973.

[3] A. O'Toole, A.J. Mistlin, A.J. Chitty, A physical system approach to recognition memory for spatially transformed faces, Neural Network 1 (1988) 179–199.

[4] M. Flemming, G. Cottrell, Categorization of faces using unsupervised feature extraction, International Joint Conference on Neural Networks, Vol. 2, 1990.

[5] T. Kohonen, P. Lehtio, Storage and processing of information in distributed associative memory systems, in: G.E. Hinton, J.A. Anderson (Eds.), Parallel Models of Associative Memory, Erlbaum, London, 1981, pp. 105–143.

[6] R. Brunelli, T. Poggio, Face recognition: features vs. templates, IEEE Trans. Pattern Anal. Mach. Intell. 15 (10) (1993) 1042–1052.

[7] M. Kirby, L. Sirovich, Application of the Karhunen–Loeve procedure for the characterization of human faces, IEEE Trans. Pattern Anal. Mach. Intell. 12 (1) (1990) 103–108.

[8] M.M. Loève, Probability Theory, Van Nostrand, Princeton, 1955.

[9] I.T. Jolliffe, Principal Component Analysis, Springer, New York, 1986.

[10] M. Turk, A. Pentland, Eigenfaces for recognition, J. Cognitive Neurosci. 3 (1) (1991).

[11] R. Chellappa, S. Sirohey, C.L. Wilson, C.S. Barnes, Human and machine recognition of faces: a survey, Technical Report CAR-TR-731, CS-TR-3339, University of Maryland, August 1994.

[12] A. Pentland, B. Moghaddam, T. Starner, View-based and modular eigenspaces for face recognition, Proceedings of IEEE Conference on Computer Vision & Pattern Recognition, Seattle, WA, June 1994.

[13] B. Moghaddam, A. Pentland, Probabilistic visual learning for object detection, IEEE Proceedings of the Fifth International Conference on Computer Vision (ICCV'95), Cambridge, USA, June 1995.

[14] B. Moghaddam, A. Pentland, Probabilistic visual learning for object representation, IEEE Trans. Pattern Anal. Mach. Intell. PAMI-19 (7) (1997) 696–710.

[15] B.J. Frey, A. Colmenarez, T.S. Huang, Mixtures of local linear subspaces for face recognition, Proceedings of IEEE

Computer Vision & Pattern Recognition (CVPR'98), June 1998, pp. 32–37.

[16] K. Etemad, R. Chellappa, Discriminant analysis for recognition of human faces, Proceedings of International Conference on Acoustics, Speech and Signal Processing, 1996, pp. 2148–2151.

[17] V.I. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. PAMI-19 (7) (1997) 711–720.

[18] D. Swets, J. Weng, Using discriminant eigenfeatures for image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. PAMI-18(8) (1996) 831–836.

[19] C. Liu, H. Wechsler, Face recognition using evolutionary pursuit, Proceedings of 5th European Conference on Computer Vision (ECCV'98), Frieburg, Germany, 1998.

[20] I. Craw, P. Cameron, Face recognition by computer, in: D. Hogg, R. Boyle (Eds.), Proceedings of the British Machine Vision Conference, Springer, Berlin, 1992, pp. 498–507.

[21] I. Craw et al., Automatic face recognition: combining configuration and texture, in: M. Bichsel (Ed.), Proceedings of the International Workshop on Automatic Face and Gesture Recognition, Zurich, 1995.

[22] A. Lanitis, C.J. Taylor, T.F. Cootes, A unified approach to coding and interpreting face images, IEEE Proceedings of the 5th International Conference on Computer Vision (ICCV'95), Cambridge, MA, June 1995.

[23] M.C. Burl, U.M. Fayyad, P. Perona, P. Smyth, M.P. Burl, Automating the hunt for volcanos on venus, Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition, Seattle, WA, June 1994.

[24] C. Bregler, S.M. Omohundro, Surface learning with applications to lip reading, Advances in Neural Information Processing Systems, Vol. 6, 1994, pp. 43–50.

[25] H. Murase, S.K. Nayar, Visual learning and recognition of 3D objects from appearance, Int. J. Comput. Vision 14 (1) (1995) 5–24.

[26] J.J. Weng, On comprehensive visual learning, Proc. NSF/ARPA Workshop on Performance vs. Methodology in Computer Vision, Seattle, WA, June 194.

[27] B. Moghaddam C. Nastar, A. Pentland, A bayesian similarity measure for direct image matching, International Conference on Pattern Recognition, Vienna, Austria, August 1996.

[28] B. Moghaddam, C. Nastar, A. Pentland, Bayseian face recognition using deformable intensity differences, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, June 1996.

[29] B. Moghaddam, W. Wahid, A. Pentland, Beyond eigenfaces: probabilistic matching for face recognition, Proceedings of the International Conference on Automatic Face and Gesture Recognition, Nara, Japan, April 1998, pp. 30–35.

[30] B. Moghaddam, T. Jebara, A. Pentland, Bayesian modeling of facial similarity, Advances in Neural Information Processing Systems, Vol. 11, 1998, pp. 910–916.

[31] B. Moghaddam, C. Nastar, A. Pentland, Bayesian face recognition using deformable intensity differences, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 1996.

[32] C. Nastar, B. Moghaddam, A. Pentland, Generalized image matching: statistical learning of physically-based deformations, Proceedings of the 4th European Conference on Computer Vision (ECCV'96), Cambridge, UK, April 1996.

[33] M.E. Tipping, C.M. Bishop, Mixtures of principal component analyzers, Proceedings of the IEEE 5th International Conference on Artificial Neural Networks, Cambridge, UK, June 1997, pp. 13–18.

[34] P.J. Phillips, H. Moon, P. Rauss, S. Rizvi, The FERET evaluation methodology for face-recognition alogrithms, IEEE Proceedings of Computer Vision and Pattern Recognition, June 1997, pp. 137–143.

[35] B. Moghaddam, A. Pentland, Face recognition using view-based and modular eigenspaces, Automatic Systems for the Indentification and Inspection of Humans, Vol. SPIE 2277, San Diego, CA, 1994, pp. 12–21.

About the Author—BABACK MOGHADDAM is a Research Scientist at Mitsubishi Electric Research Laboratory in Cambridge MA, USA. He received the B.S. (Magna Cum Laude) and M.S. (Honors) degrees in Electrical & Computer Engineering from George Mason University in 1989 and 1992, respectively, and his Ph.D. degree from the Department of Electrical Engineering and Computer Science at the Massachusettes Institute of Technology in 1997. During his doctoral studies he was a Research Assistant in the Vision & Modeling group at the MIT Media Laboratory, where he developed an automatic face recognition system that was the top competitor ion ARPA's "FERET" face recognition competition. His research interests include computer vision, image processing, computational learning theory and statistical pattern recognition. He is a member of Eta Kappa Nu, IEEE and ACM.

About the Author—TONY JEBARA received his BEng in honours electrical engineering from McGill University, Canada in 1996. He also worked at the McGill Center for Intelligent Machines from 1994–1996 on computer vision and 3D face recognition research. In 1996, he joined the MIT Media Laboratory to work at the Vision and Modeling — Perceptual Computing Group. In 1998, he obtained his M.Sc. degree in Media Arts and Sciences for research in real-time 3D tracking and visual interactive behavior modeling. He is currently pursuing a Ph.D. degree at the MIT Media Laboratory and his interests include computer vision, machine learning, wearable computing and behavior modeling.

About the Author—ALEX PENTLAND is the Academic Head of the M.I.T. Media Laboratory. He is also the Toshiba Professor of Media Arts and Sciences, and endowed chair last held by Marvin Minsky. He received his Ph.D. from the Massachusetts Institute of Technology in 1982. He then worked at SRI's AI Center and as a Lecturer at Stanford University, winning the Distinguished Lecturer award in 1986. In 1987 he returned to M.I.T. to found the Perceptual Computing Section of the Media Laboratory, a group that now includes over 50 researches in computer vision, graphics, speech, music, and human–machine interaction. He has done reseach in human–machine interface, computer graphics, artifical intelligence, machine and human vision, and has published more than 180 scientific articles in these areas. His most recent research focus is understanding human behavior in video, including face, expression, gesture, and intention recognition, as described in the April 1996 issue of Scientific American. He has won awards from the AAAI for his research into fractals; the IEEE for his research into face recognition; and from Ars Electronica for his work in computer vision interfaces to virtual environments.