M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 443
Appears in: *The 3rd IEEE Int'l Conference on Automatic Face & Gesture Recognition, Nara, Japan, April 1998*

Beyond Eigenfaces: Probabilistic Matching for Face Recognition

Baback Moghaddam, Wasiuddin Wahid and Alex Pentland

MIT Media Laboratory, 20 Ames St., Cambridge, MA 02139, USA.
Email: {baback,wasi,sandy}@media.mit.edu

Abstract

We propose a novel technique for direct visual matching of images for the purposes of face recognition and database search. Specifically, we argue in favor of a *probabilistic* measure of similarity, in contrast to simpler methods which are based on standard L_2 norms (*e.g.,* template matching) or subspace-restricted norms (*e.g.,* eigenspace matching). The proposed similarity measure is based on a Bayesian analysis of image differences: we model two mutually exclusive classes of variation between two facial images: *intra-personal* (variations in appearance of the same individual, due to different expressions or lighting) and *extra-personal* (variations in appearance due to a difference in identity). The high-dimensional probability density functions for each respective class are then obtained from training data using an 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, which is used to rank matches in the database. The performance advantage of this probabilistic matching technique over standard nearest-neighbor eigenspace matching is demonstrated using results from ARPA's 1996 "FERET" face recognition competition, in which this algorithm was found to be the top performer.

1 Introduction

Current approaches to image matching for visual object recognition and image database retrieval often make use of simple image similarity metrics such as Euclidean distance or normalized correlation, which correspond to a standard template-matching approach to recognition [2]. For example, in its simplest form, the similarity measure $S(I_1, I_2)$ between two images I_1 and I_2 can be set to be inversely proportional to the norm $||I_1 - I_2||$. Such a simple formulation suffers from a major drawback: it does not exploit knowledge of which type of variations are critical (as opposed to incidental) in expressing similarity. In this paper, we formulate a *probabilistic* similarity measure which is based on the probability that the image intensity differences, denoted by $\Delta = I_1 - I_2$, are characteristic of typical variations in appearance of the *same* object. For example, for purposes of face recognition, we can 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

$$S(I_1, I_2) \;=\; P(\Delta \in \Omega_I) \;=\; P(\Omega_I|\Delta) \qquad (1)$$

where $P(\Omega_I|\Delta)$ is the *a posteriori* probability given by Bayes rule, using estimates of the likelihoods $P(\Delta|\Omega_I)$ and $P(\Delta|\Omega_E)$ which are derived from training data using an efficient subspace method for density estimation of high-dimensional data [6]. This Bayesian (MAP) approach can also be viewed as a generalized nonlinear extension of Linear Discriminant Analysis (LDA) [8, 3] or "FisherFace" techniques [1] for face recognition. Moreover, our nonlinear generalization has distinct computational/storage advantages over these linear methods for large databases.

2 Analysis of Intensity Differences

We now consider the problem of characterizing the type of differences which occur when matching two images in a face recognition task. 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 which result when 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 define the 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:

$$\begin{aligned} S &= P(\Omega_I|\Delta) \\ &= \frac{P(\Delta|\Omega_I)P(\Omega_I)}{P(\Delta|\Omega_I)P(\Omega_I) + P(\Delta|\Omega_E)P(\Omega_E)} \end{aligned} \qquad (2)$$

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. Additionally, 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 much 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 $S(I_1, I_2) > \frac{1}{2}$.

2.1 Density Modeling

One difficulty with this approach is that the intensity difference vector is very high-dimensional, with $\Delta \in \mathcal{R}^N$ and $N = O(10^4)$. Therefore we typically lack sufficient independent training observations to compute reliable 2nd-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 for each class is likely to be significantly smaller than N.

Recently, an efficient density estimation method was proposed by Moghaddam & Pentland [6] which divides the vector space $\mathcal{R}^N$ into two complementary subspaces using an eigenspace decomposition. This method relies on a Principal Components Analysis (PCA) [4] to form a low-dimensional estimate of the complete likelihood which can be evaluated using only the first M principal components, where $M << N$. This decomposition is illustrated in Figure 1 which shows an orthogonal decomposition of the vector space $\mathcal{R}^N$ into two mutually exclusive subspaces: the principal subspace F containing the first M principal components and its orthogonal complement $\bar{F}$, which contains the residual of the expansion. The component in the orthogonal subspace $\bar{F}$ 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.

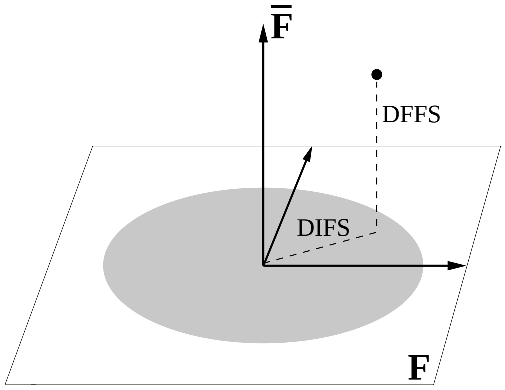
As shown in [6], the complete likelihood estimate can be written as the product of two independent marginal Gaussian densities

$$\begin{aligned} \hat{P}(\Delta|\Omega) &= \left[\frac{\exp\left(-\frac{1}{2}\sum_{i=1}^{M}\frac{y_i^2}{\lambda_i}\right)}{(2\pi)^{M/2}\prod_{i=1}^{M}\lambda_i^{1/2}} \right] \cdot \left[\frac{\exp\left(-\frac{\epsilon^2(\Delta)}{2\rho}\right)}{(2\pi\rho)^{(N-M)/2}} \right] \\ &= P_F(\Delta|\Omega)\ \hat{P}_{\bar{F}}(\Delta|\Omega) \end{aligned} \tag{3}$$

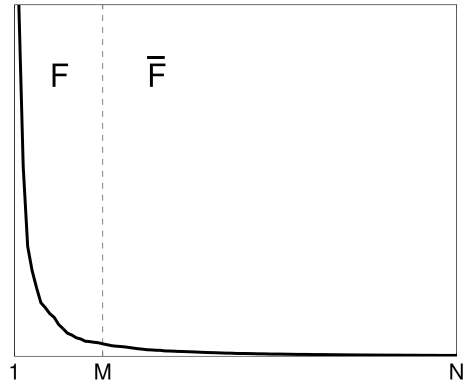
where $P_F(\Delta|\Omega)$ is the true marginal density in F, $\hat{P}_{\bar{F}}(\Delta|\Omega)$ is the estimated marginal density in the orthogonal complement $\bar{F}$, y_i are the principal components and $\epsilon^2(\Delta)$ is the residual (or DFFS). The optimal value for the weighting parameter ρ is then found to be simply the average of the $\bar{F}$ eigenvalues

$$\rho = \frac{1}{N-M}\sum_{i=M+1}^{N}\lambda_i \tag{4}$$

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.



(a)



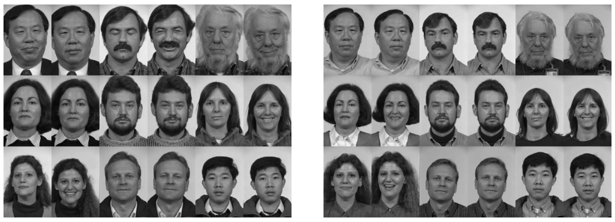
(b)

Figure 1: (a) Decomposition of $\mathcal{R}^N$ 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.

3 Experiments

To test our recognition strategy we used a collection of images from the FERET face database. This collection of images consists of hard recognition cases that have proven difficult for all face recognition algorithms previously tested on the FERET database. The difficulty posed by this dataset 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) and 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 Figure 2).

Before we can apply our matching technique, we need to perform an affine alignment of these facial images. For this purpose we have used an automatic face-processing system 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 [6] 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 Figure 3.



(a) (b)

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

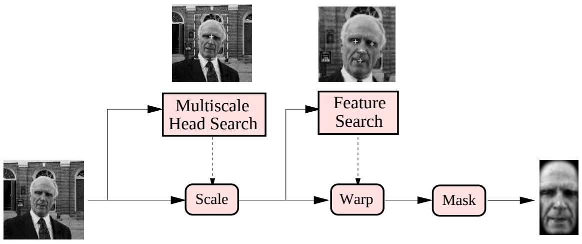


Figure 3: The face alignment system

All the faces in our experiments were geometrically aligned and normalized in this manner prior to further analysis.

3.1 Eigenface Matching

As a baseline comparison, we first used an eigenface matching technique for recognition [9]. The normalized images from the gallery and the probe sets were projected onto a 100-dimensional eigenspace and a nearest-neighbor rule based on a Euclidean distance measure was used to match each probe image to a gallery image. We note that this method corresponds to a generalized template-matching method which uses a Euclidean norm type of similarity $S(I_1, I_2)$, which is restricted to the principal component subspace of the data. A few of the lower-order eigenfaces used for this projection are shown in Figure 4. We note 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. Note that this performance is better than or similar to recognition rates obtained by any algorithm tested on this database, and that it is lower (by about 10%) than the typical rates that we have obtained with the FERET database [5]. We attribute this lower performance to the fact that these images were selected to be particularly challenging. In fact, using an eigenface method to match the first views of the 76 individuals in the gallery to their second views, we obtain a higher recognition rate of 89% (68 out of 76), suggesting that the gallery images represent a less challenging data set since these images were taken at the same time and under identical lighting conditions.

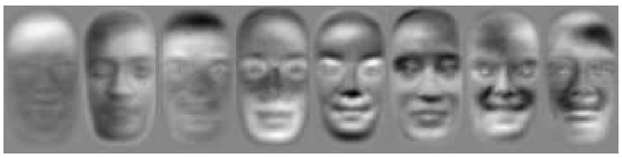


Figure 4: Standard Eigenfaces.

3.2 Bayesian Matching

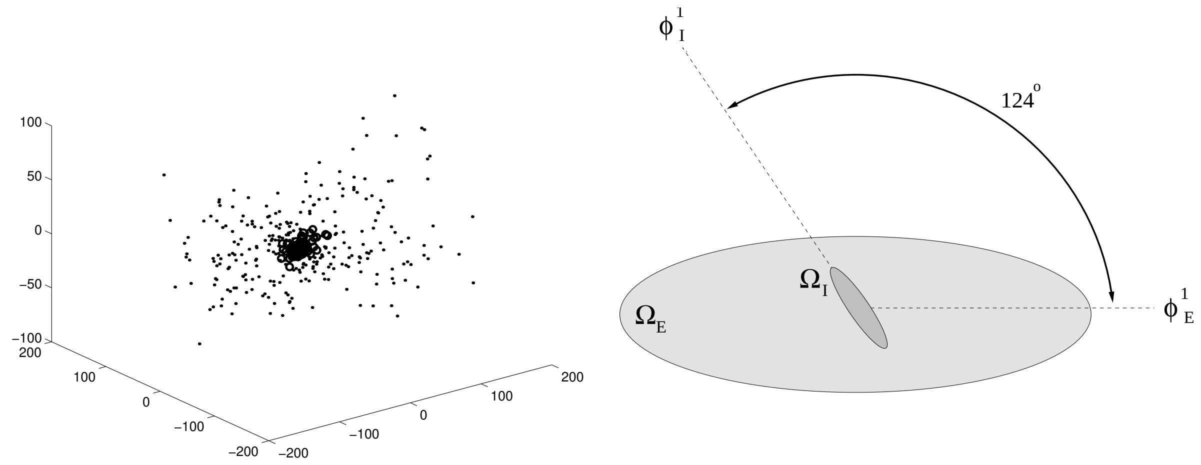
For our probabilistic algorithm, we first gathered training data by computing the intensity differences for a training subset of 74 intrapersonal differences (by matching the two views of every individual in the gallery) and a random subset of 296 extrapersonal differences (by matching images of *different*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

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 each vector onto the principal eigenvectors. Such a visualization is shown in Figure 5(a) which is a 3D scatter plot of the first 3 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 can not reliably distinguish low-amplitude extrapersonal differences (of which there are many) from intrapersonal ones.

However, direct visual interpretation of Figure 5(a) is very 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. Fortunately, 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°, implying that the orientation of the two hyper-ellipsoids is quite different. Figure 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.

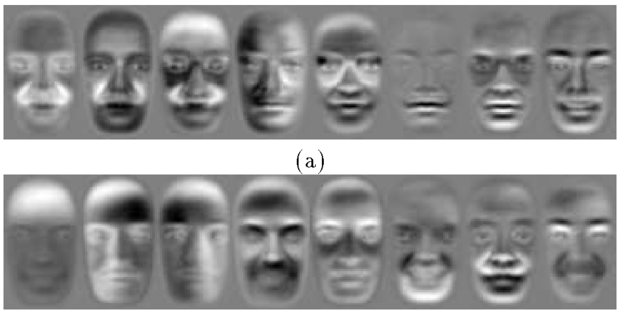
3.3 Dual Eigenfaces

We note that the two mutually exclusive classes Ω_I and Ω_E correspond to a "dual" set of eigenfaces as shown in Figure 6. Note that the intrapersonal variations shown in Figure 6-(a) represent subtle variations due mostly to expression changes (and lighting) whereas the extrapersonal variations in Figure 6-(b) are more representative of general eigenfaces which code variations such as hair color, facial hair and glasses. Also note the overall qualitative similarity of the extrapersonal eigenfaces to the standard eigenfaces in Figure 4. This suggests the basic intuition that intensity



(a) (b)

Figure 5: (a) Distribution of the two classes in the first 3 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.



(a)

(b)

Figure 6: "Dual" Eigenfaces: (a) Intrapersonal, (b) Extrapersonal

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 part of formulating a probabilistic measure of facial similarity. In fact our experiments with a larger set of FERET images have shown that this intrapersonal eigenspace alone is sufficient for a simplified *maximum likelihood* measure of similarity (see Section 3.4).

Finally, we note that since these classes are not linearly separable, simple linear discriminant techniques (*e.g.,* using hyperplanes) can not be used with any degree of reliability. The proper decision surface is inherently nonlinear (quadratic, in fact, 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 analyzed the geometry of the two distributions, we then computed the likelihood estimates $P(\Delta|\Omega_I)$ and $P(\Delta|\Omega_E)$ using the PCA-based method outlined in Section 2.1. We selected 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 $P(\Omega_I|\Delta)$ for matching probe images to those in the gallery.

Therefore, for each probe image we computed probe-to-gallery differences and sorted the matching order, this time using the *a posteriori* probability $P(\Omega_I|\Delta)$ as the similarity measure. 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 $P(\Omega_I|\Delta) > P(\Omega_E|\Delta)$.

3.4 The 1996 FERET Competition Results

This approach to recognition has produced a significant improvement over the accuracy we obtained using 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 [7] (see Figure 7). Figure 8 shows the performance comparison between standard eigenfaces and the Bayesian method from this test. Note the 10% gain in performance afforded by the new Bayesian similarity measure. Similarly, Figure 9 shows the recognition results for "duplicate" images which were separated in time by up to 6 months (a much more challenging recognition problem) which shows a 30% improvement in recognition rate with Bayesian matching. Thus we note that in both cases (FA/FB and duplicates) the new probabilistic similarity measure has effectively *halved* the error rate of eigenface matching.

We have recently experimented with a more simplified probabilistic similarity measure which uses only the *in-*

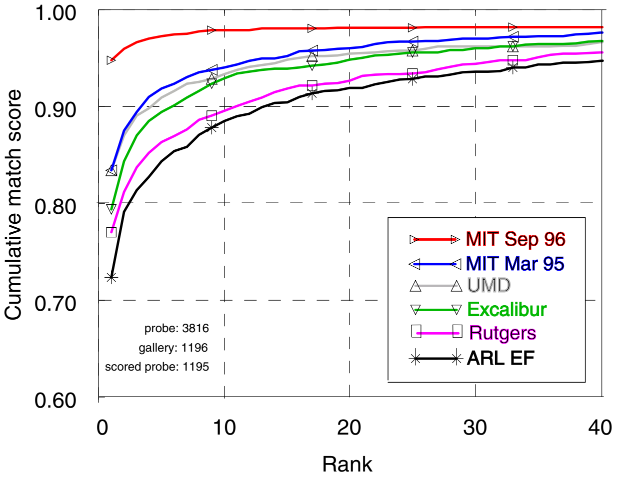


Figure 7: Cumulative recognition rates for frontal FA/FB views for the competing algorithms in the FERET 1996 test. The top curve (labeled "MIT Sep 96") corresponds to our Bayesian matching technique. Note that second placed is standard eigenface matching (labeled "MIT Mar 95").

trapersonal eigenfaces with the intensity difference Δ to formulate a *maximum likelihood* (ML) matching technique using

$$S' = P(\Delta|\Omega_I) \tag{5}$$

instead of the *maximum a posteriori* (MAP) approach defined by Equation 2. Although this simplified measure has not yet been officially FERET tested, our own experiments with a database of size 2000 have shown that using S' instead of S results in only a minor (2%) deficit in the recognition rate while cutting the computational cost by a factor of 1/2 (requiring a single eigenspace projection as opposed to two).

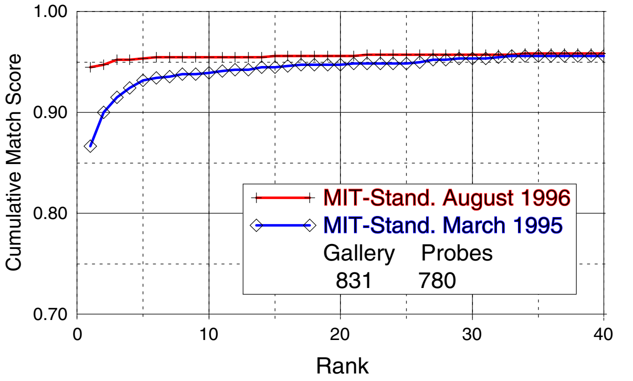


Figure 8: Cumulative recognition rates for frontal FA/FB views with standard eigenface matching and the newer Bayesian similarity metric.

4 Conclusions

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

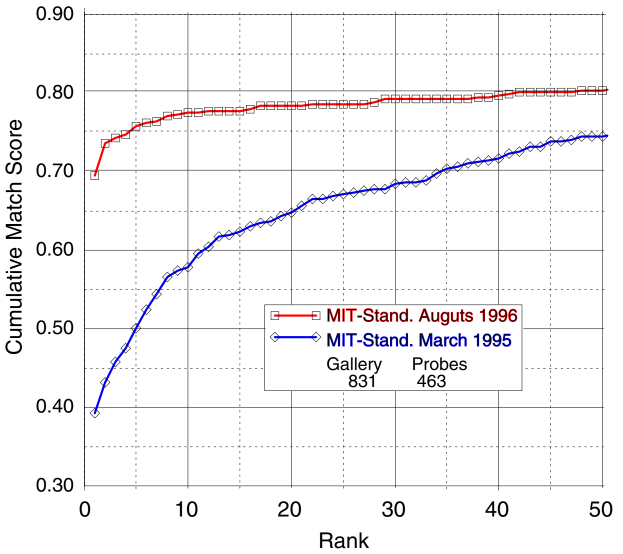


Figure 9: Cumulative recognition rates for frontal duplicate views with standard eigenface matching and the newer Bayesian similarity metric.

to simpler methods which are based on standard L_2 norms (*e.g.,* template matching [2]) or subspace-restricted norms (*e.g.,* eigenspace matching [9]). This technique is based on a Bayesian analysis of image differences which leads to a very useful measure of similarity.

The performance advantage of our probabilistic matching technique has been demonstrated using both a small database (internally tested) as well as a large (800+) 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 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 deformations 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. Furthermore, by equating similarity with the *a posteriori* probability we obtain an optimal non-linear decision rule for matching and recognition. This aspect of our approach differs significantly from recent methods which use simple linear discriminant analysis techniques for recognition (*e.g.,* [8, 3]). Our Bayesian (MAP) method can also be viewed as a generalized nonlinear (quadratic) version of Linear Discriminant Analysis (LDA) [3] or "FisherFace" techniques [1]. 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 LDA). One (or at most two) eigenspaces are sufficient for probabilistic matching and therefore storage and computational costs are fixed and do not increase with the size of the database (as with LDA/Fisher methods).

Finally, the results obtained with the simplified ML similarity measure (S' in Eq. 5) 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. For completeness (and a slightly better performance) however, one should use the *a posteriori* similarity S in Eq. 2, at twice the computational cost of standard eigenfaces.

References

[1] V.I. Belhumeur, J.P. Hespanha, and D.J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-19(7):711–720, July 1997.

[2] R. Brunelli and T. Poggio. Face recognition : Features vs. templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(10), October 1993.

[3] K. Etemad and R. Chellappa. Discriminant analysis for recognition of human faces. In *Proc. of Int'l Conf. on Acoustics, Speech and Signal Processing*, pages 2148–2151, 1996.

[4] I.T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, 1986.

[5] B. Moghaddam and A. Pentland. Face recognition using view-based and modular eigenspaces. *Automatic Systems for the Identification and Inspection of Humans*, 2277, 1994.

[6] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-19(7):696–710, July 1997.

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

[8] D. Swets and J. Weng. Using discriminant eigenfeatures for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-18(8):831–836, August 1996.

[9] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1), 1991.