M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 393
Appears in: *International Conference on Pattern Recognition, Vienna, Austria, August 1996.*

A Bayesian Similarity Measure for Direct Image Matching

Baback Moghaddam, Chahab Nastar and Alex Pentland
Perceptual Computing Section, The Media Laboratory,
Massachusetts Institute of Technology
20 Ames Street, Cambridge MA 02139, U.S.A.

Abstract

We propose a probabilistic similarity measure for direct image matching based on a Bayesian analysis of image deformations. We model two classes of variation in object appearance: intra-object *and* extra-object*. The probability density functions for each class are then estimated from training data and used to compute a similarity measure based on the* a posteriori *probabilities. Furthermore, we use a novel representation for characterizing image differences using a deformable technique for obtaining pixel-wise correspondences. This representation, which is based on a deformable 3D mesh in XYI-space, is then experimentally compared with two simpler representations: intensity differences and optical flow. The performance advantage of our deformable matching technique is demonstrated using a typically hard test set drawn from the US Army's FERET face database.*

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. 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 $||I_2 - I_1||$. Such a simple formulation suffers from two major drawbacks: it requires precise alignment of the objects in the image and 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-based differences, denoted by $d(I_1, I_2)$, 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

$$S(I_1, I_2) \;=\; P(d(I_1, I_2) \in \Omega_I) \;=\; P(\Omega_I | d(I_1, I_2)) \quad (1)$$

where $P(\Omega_I | d(I_1, I_2))$ is the *a posteriori* probability given by Bayes rule, using estimates of the likelihoods $P(d(I_1, I_2)|\Omega_I)$ and $P(d(I_1, I_2)|\Omega_E)$ which are derived from training data using an efficient subspace method for density estimation of high-dimensional data [9].

Furthermore, we use a novel representation for $d(I_1, I_2)$ which combines both the spatial (XY) and grayscale (I) components of the image in a unified XYI framework (unlike previous approaches which essentially treat the shape and texture components independently, *e.g.,* [3, 4, 7, 2]). Specifically, I_1 is modeled as a physically-based deformable 3D surface (or manifold) in XYI-space which deforms in accordance with attractive "physical forces" exerted by I_2. The dynamics of this system are efficiently solved for using the *analytic modes of vibration* [10], yielding a 3D correspondence field for warping I_1 into I_2. In addition, we use the *parametric* representation, $d(I_1, I_2) = \tilde{\mathbf{U}}$, where $\tilde{\mathbf{U}}$ is the modal amplitude spectrum of the resultant deformation [12]. This manifold matching technique can be viewed as a more general formulation for image correspondence which, unlike optical flow, does *not* require a constant brightness assumption [5].[1]

Finally, we experimentally compare our deformable matching technique with two alternative (non-deformable) methods: one using intensity differences with $d(I_1, I_2) = I_2 - I_1$, and a standard correspondence method using optical flow with $d(I_1, I_2) = flow(I_1, I_2)$ where $flow(I_1, I_2)$ is the vector flow field between I_1 and I_2. We note that these two methods can be viewed as degenerate cases of our general XYI correspondence method: the former assumes XY correspondences and makes the I difference explicit, whereas the latter assumes comparable I components and makes the XY variations explicit. Our experimental results have confirmed our basic intuition that the fully deformable XYI warping method yields the best characterization of $d(I_1, I_2)$, at least as far as recognition is concerned. The advantage of our method over optical flow is key, since this simpler method relies all too heavily on the constant brightness assumption and is prone to failure when there are large grayscale variations between the images of different individuals (*e.g.,* presence/absence of facial hair).

2 XYI Image Warping

In previous work [12], we formulated an image matching technique based on a 3D surface representation of an image $I(x, y)$ — *i.e.,* as the surface $(x, y, I(x, y))$ as shown, for example, in Figure 1 — and developed an efficient method to *warp* one image onto another using a physically-based deformation model. In this section we briefly review the mathematics of this approach (for further details the reader

[1] In fact, by simply disabling the I component of our deformations we can obtain a standard 2D deformable mesh which yields correspondences similar to an optical flow technique with thin-plate regularizers.

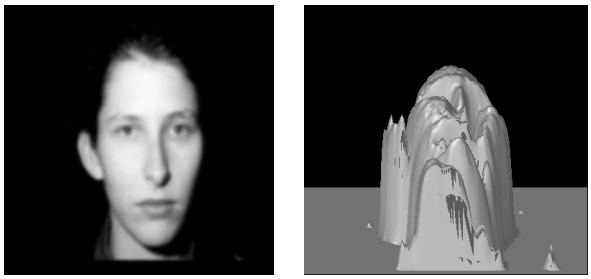


Figure 1: An image and its XYI surface representation

is referred to [11, 12]).

The intensity surface is modeled as a deformable mesh and is governed by Lagrangian dynamics [1] :

$$\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{F}(t) \tag{2}$$

where $\mathbf{U} = [\ldots, \Delta x_i, \Delta y_i, \Delta z_i, \ldots]^T$ is a vector storing nodal displacements, $\mathbf{M}$, $\mathbf{C}$ and $\mathbf{K}$ are respectively the mass, damping and stiffness matrices of the system, and $\mathbf{F}$ is the external force. In warping one image onto a second (reference) image, the external force at each node M_i of the mesh points is the vector to the closest 3D point P_i in the reference surface:

$$\mathbf{F}(t) = [\ldots, \overrightarrow{M_i P_i}(t), \ldots]^T \tag{3}$$

The final correspondence (and consequently the resultant XYI-warp) between two images is obtained by solving the governing equation above. Figure 2 shows a schematic representation of the deformation process. Note that the external forces (dashed arrows) do *not* necessarily correspond to the final displacement field of the surface. The elasticity of the surface provides an intrinsic smoothness constraint for computing the final displacement field.

We note that this formulation provides an interesting alternative to optical flow methods for obtaining correspondence, without the classical *brightness constraint* [5]. Indeed, the brightness constraint corresponds to a particular case of our formulation where the closest point P_i has to have the same intensity as M_i — *i.e.,* $\overrightarrow{M_i P_i}$ is parallel to the XY plane. We do not make that assumption here.

Solutions of the governing equation are typically obtained using an eigenvector-based *modal* decomposition [13, 11, 10]. In particular, the vibration modes $\phi(i)$ of the previous deformable surface are the vector solutions of the eigenproblem :

$$\mathbf{K}\phi = \omega^2 \mathbf{M}\phi \tag{4}$$

where $\omega(i)$ is the i-th eigenfrequency of the system. Solving the governing equations in the modal basis leads to scalar equations where the unknown $\tilde{u}(i)$ is the amplitude of mode i [1]

$$\ddot{\tilde{u}}(i) + \tilde{c}_i \dot{\tilde{u}}(i) + \omega(i)^2 \tilde{u}(i) = \tilde{f}_i(t) \qquad i = 1, \ldots, 3N. \tag{5}$$

The closed-form expression of the displacement field is then given by

$$\mathbf{U} \approx \sum_{i=1}^{P} \tilde{u}(i)\phi(i) \tag{6}$$

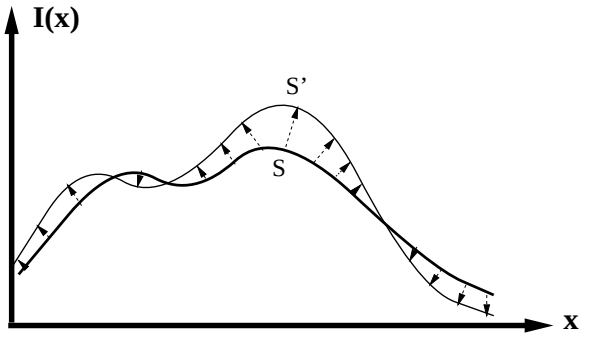


Figure 2: A cross-section of the intensity surface S being pulled towards S' by image forces

with $P \ll 3N$, which means that only P scalar equations of the type of (5) need to be solved. The modal superposition equation (6) can be seen as a Fourier expansion with high-frequencies neglected [10]. In our formulation, however, we make use of the *analytic modes* [10, 12], which are known sine and cosine functions for specific surface topologies

$$\phi(p, p') = [\ldots, \cos\frac{p\pi(2i-1)}{2n} \cos\frac{p'\pi(2j-1)}{2n'}, \ldots]^T \tag{7}$$

These analytic expressions avoid costly eigenvector decompositions and furthermore allow the total number of modes to be easily adjusted for the application.

The above modal analysis technique represents a coordinate transform from the nodal displacement space to the modal amplitude subspace:

$$\tilde{\mathbf{U}} = \mathbf{\Phi}^T \mathbf{U} \tag{8}$$

where $\mathbf{\Phi}$ is the matrix of analytic modes $\phi(p, p')$ and $\tilde{\mathbf{U}}$ is the resultant vector of modal amplitudes which encodes the type of deformations which characterize the difference between the two images. In addition, once we have solved for the resultant 3D displacement field we can then warp the original image onto the second in the XYI space and then render a resultant 2D image using simple computer graphics techniques. Figure 3 shows an example illustrating this warping process. We note that the warped image $I_{1\to 2}$ is only an incidental by-product of our correspondence method. Since our main goal is image matching we are primarily interested in the modal amplitude spectrum $\tilde{\mathbf{U}}$ for expressing $d(I_1, I_2)$.

3 Analysis of Deformations

We now consider the problem of characterizing the type of deformations 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 $P(\tilde{\mathbf{U}}|\Omega_I)$ and $P(\tilde{\mathbf{U}}|\Omega_E)$ for a given deformation's modal amplitude vector $\tilde{\mathbf{U}}$.

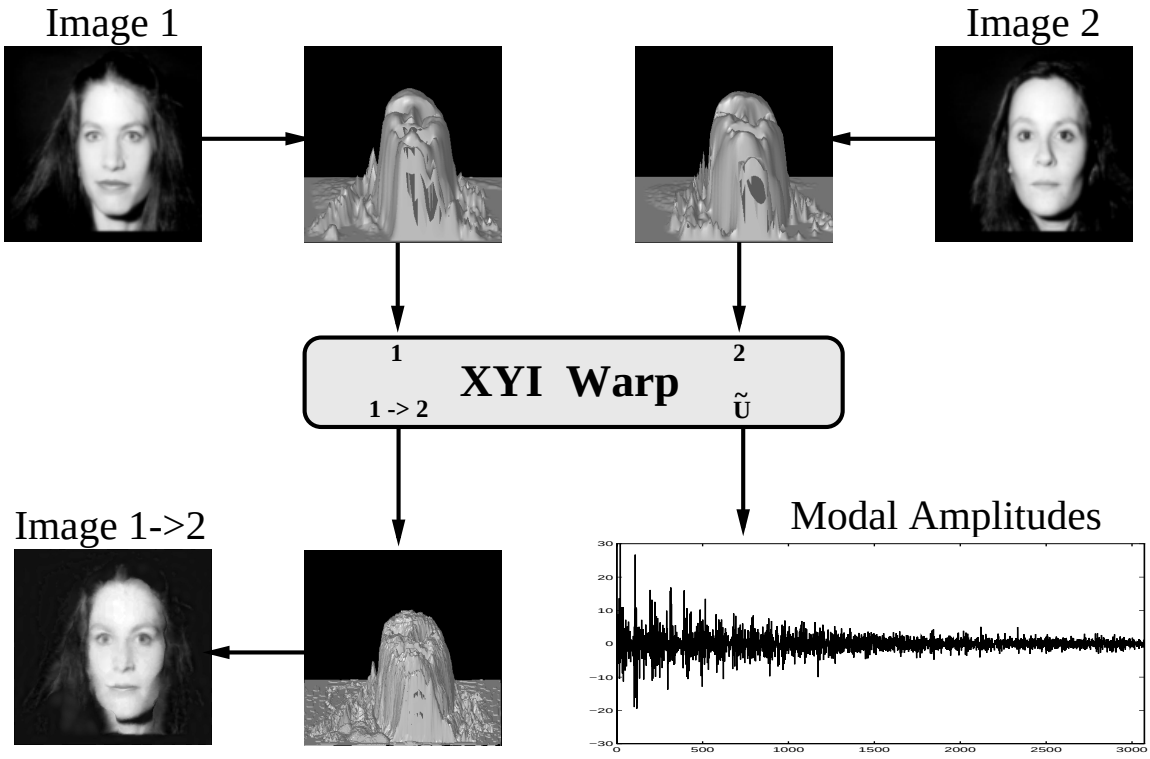


Figure 3: Example of XYI warping two images.

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:

$$S(I_1, I_2) = P(\Omega_I|\tilde{\mathbf{U}}) = \frac{P(\tilde{\mathbf{U}}|\Omega_I)P(\Omega_I)}{P(\tilde{\mathbf{U}}|\Omega_I)P(\Omega_I) + P(\tilde{\mathbf{U}}|\Omega_E)P(\Omega_E)} \quad (9)$$

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 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 $P(\Omega_I|\tilde{\mathbf{U}}) > P(\Omega_E|\tilde{\mathbf{U}})$, or equivalently, if $S(I_1, I_2) > \frac{1}{2}$.

3.1 Statistical Modeling of Modes

One difficulty with this approach is that the modal amplitude vectors are high-dimensional, with $\tilde{\mathbf{U}} \in \mathcal{R}^N$ with $N = O(10^3)$. 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 $\tilde{\mathbf{U}}$ for each class is likely to be significantly smaller than N.

Recently, an efficient density estimation method was proposed by Moghaddam & Pentland [9] 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) [6] 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 4 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 $\tilde{\mathbf{U}}$ 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 [9], the complete likelihood estimate can be written as the product of two independent marginal

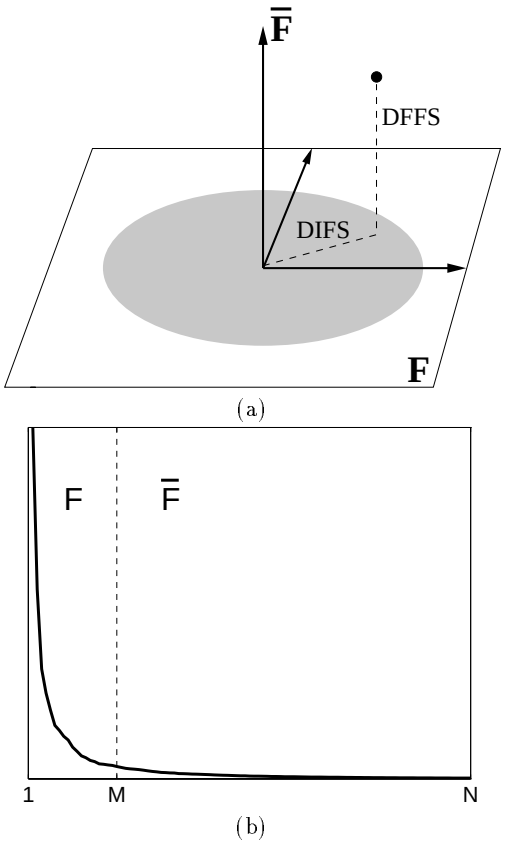


Figure 4: (a) Decomposition of $\mathcal{R}^N$ into the principal subspace F and its orthogonal complement $\bar{F}$ for a Gaussian density, (b) a typical eigenvalue spectrum and its division into the two orthogonal subspaces.

Gaussian densities

$$\hat{P}(\tilde{\mathbf{U}}|\Omega) = \left[\frac{\exp\left(-\frac{1}{2}\sum_{i=1}^{M}\frac{y_i^2}{\lambda_i}\right)}{(2\pi)^{M/2}\prod_{i=1}^{M}\lambda_i^{1/2}}\right]\cdot\left[\frac{\exp\left(-\frac{\epsilon^2(\tilde{\mathbf{U}})}{2\rho}\right)}{(2\pi\rho)^{(N-M)/2}}\right]$$
$$= P_F(\tilde{\mathbf{U}}|\Omega)\ \hat{P}_{\bar{F}}(\tilde{\mathbf{U}}|\Omega) \quad (10)$$

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

$$\rho = \frac{1}{N-M}\sum_{i=M+1}^{N}\lambda_i \quad (11)$$

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.

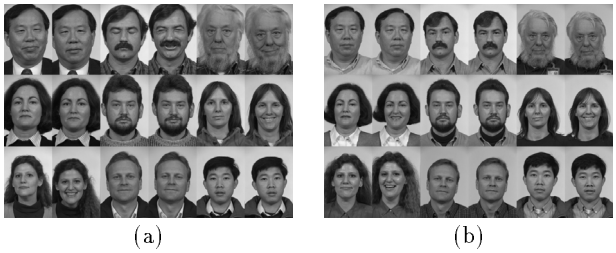


Figure 5: Examples of FERET frontal-view image pairs used for (a) the Gallery set (training) and (b) the Probe set (testing).

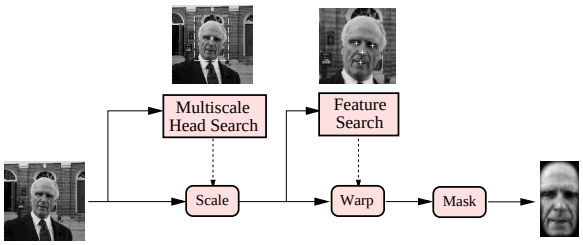


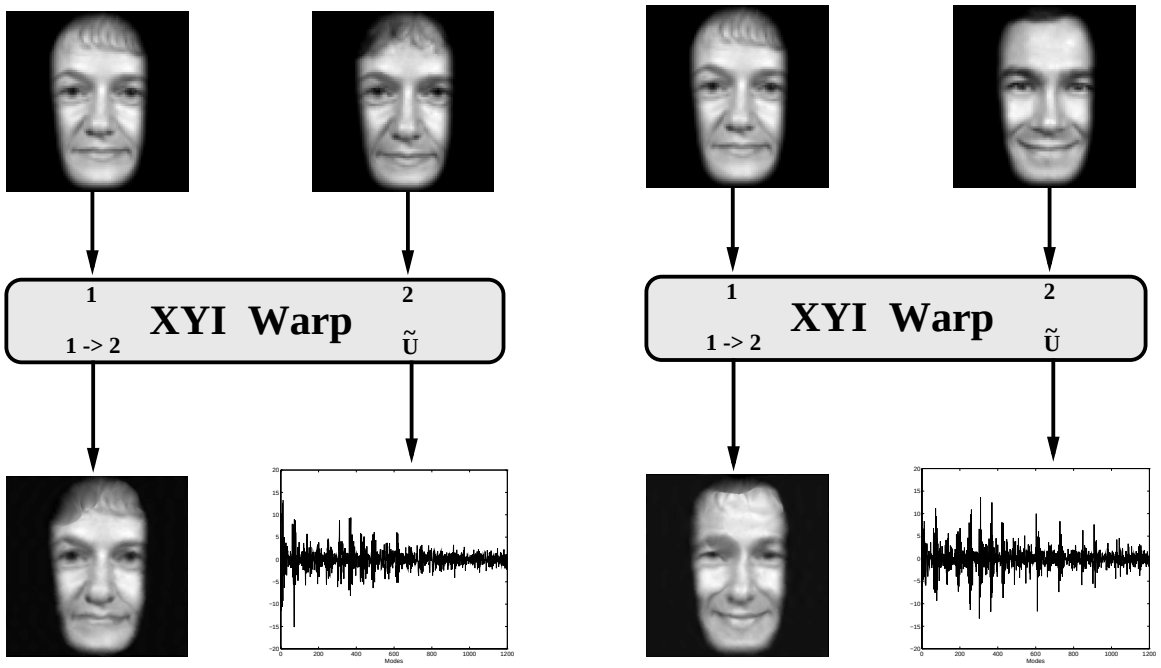
Figure 6: The face alignment system

4 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 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. These images are shown in Figure 5.

Before we can apply our deformable matching technique, we need to perform a rigid 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 Moghaddam & Pentland [9] 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 6. All the faces in our experiments were geometrically aligned and normalized in

Figure 7: The first 8 normalized eigenfaces.



(a) (b)

Figure 8: Examples of (a) intrapersonal and (b) extrapersonal facial warps.

this manner prior to further analysis.

4.1 Matching with Eigenfaces

As a baseline comparison, we first used an eigenface matching technique for recognition. 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.[2] A few of the lower-order eigenfaces used for this projection are shown in Figure 7. 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 [8]. 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.

[2]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.

4.2 Matching with XYI Deformations

For our probabilistic algorithm, we first gathered training data by computing the modal amplitude spectra for a training subset of 74 intrapersonal warps (by matching the two views of every individual in the gallery) and a random subset of 296 extrapersonal warps (by matching images of *different* individuals in the gallery), corresponding to the classes Ω_I and Ω_E, respectively. An example of each of these two types of warps is shown in Figure 8.

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 9(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 modal amplitudes as expected. It therefore appears that one can not reliably distinguish low-amplitude extrapersonal warps (of which there are many) from intrapersonal ones.

However, direct visual interpretation of Figure 9(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

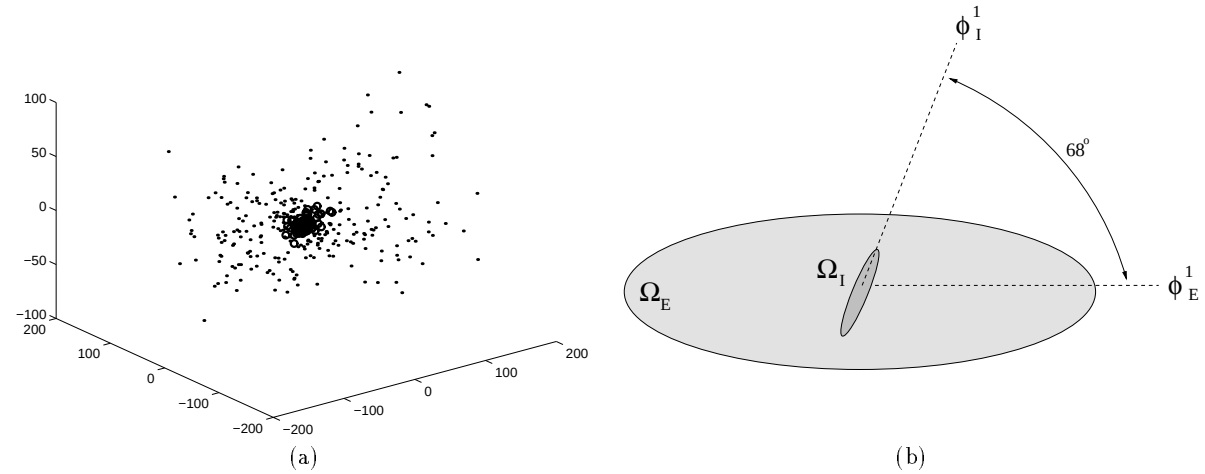


Figure 9: (a) distribution of the two classes in the first 3 principal components (circles for Ω_I, dots for Ω_E) and (b) schematic representation of the two distributions showing orientation difference between the corresponding principal eigenvectors.

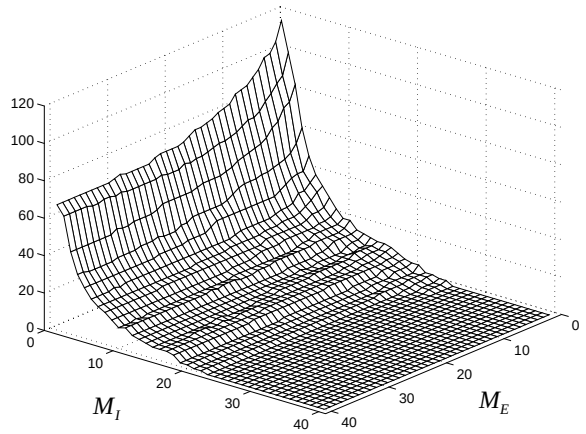


Figure 10: Total number of misclassified extrapersonal matches (with $P(\Omega_I|\tilde{\mathbf{U}}) > 0.5$) as a function of the principal subspace dimensionalities M_I and M_E.

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 68°, implying that the orientation of the two hyper-ellipsoids is quite different. Figure 9(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.

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 $P(\Omega_I|\tilde{\mathbf{U}}) = P(\Omega_E|\tilde{\mathbf{U}})$. 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 $P(\tilde{\mathbf{U}}|\Omega_I)$ and $P(\tilde{\mathbf{U}}|\Omega_E)$ using the PCA-based method outlined in Section 3.1. 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 $P(\Omega_I|\tilde{\mathbf{U}})$ for matching probe images to those in the gallery.

In order to avoid an unnecessarily large number of XYI warps, we only matched a probe image to the top 10 gallery images retrieved by the eigenface method. This significantly reduces the computational cost of our system, since computing eigenface similarity scores is negligible compared to computing XYI warps (the former takes several milliseconds whereas the latter takes approximately 20 seconds on an HP 735 workstation).

Therefore, for each probe image we computed a set of 10 probe-to-gallery warps and re-sorted the matching order, this time using the *a posteriori* probability $P(\Omega_I|\tilde{\mathbf{U}})$ as the similarity measure. This probabilistic ranking yielded an improved rank-1 recognition rate of 92% (70 out of 76). Furthermore, out of the 608 extrapersonal warps performed in this recognition experiment, only 2% (11) were misclassified as being intrapersonal — *i.e.,* with $P(\Omega_I|\tilde{\mathbf{U}}) > P(\Omega_E|\tilde{\mathbf{U}})$.

We also analyzed the sensitivity of our Bayesian matching technique with respect to the principal subspace dimensionalities M_I and M_E, which are used in estimating the likelihoods $P(\tilde{\mathbf{U}}|\Omega_I)$ and $P(\tilde{\mathbf{U}}|\Omega_E)$. The higher we set these parameters the more accurate an estimate of the likelihoods we obtain, while also requiring more principal projections. These parameters therefore represent an accuracy *vs.* complexity tradeoff in our Bayesian approach. To quantify this tradeoff, we repeated the probe set recognition experiment while varying both parameters and noted that the recognition rate never dropped below 79%, even when the two subspaces used in estimating the likelihoods were as low as one-dimensional. However, we noted that the total number of extrapersonal matches which were

	XYI-warp	I-diff	XY-flow
Mean Correct Recognition Rate	86.8 %	85.9 %	82.3 %
Max Correct Recognition Rate	92.1 %	89.5 %	86.8 %
Mean Number of False Matches	10	14	1
Max Number of False Matches	115	155	53

Figure 11: Performance of Bayesian classifier with three different data representations: full XYI-warp, intensity differences (I-diff) and optical flow (XY-flow). Results are mean/maximum values over nearly 2000 experimental trials with varying M_I and M_E.

misclassified as being intrapersonal — *i.e.,* $P(\Omega_I|\tilde{\mathbf{U}}) > P(\Omega_E|\tilde{\mathbf{U}})$ — varied in a principled way with the subspace dimensionalities. This variation is shown in Figure 10 and is clearly the type of behavior one would expect: the total number of misclassified matches decreases with increasing subspace dimensionalities. From the figure, it is apparent that these errors are more sensitively dependent on M_I, the dimensionality of the intrapersonal subspace (possibly because this class has a much lower *intrinsic* dimensionality and its distribution can be modeled using fewer principal eigenvectors).

4.3 Matching with Optical Flow and Intensity Differences

To compare the efficacy of our deformable representation for $d(I_1, I_2)$ (*i.e.,* the modal amplitudes of an XYI-warp), we next applied our Bayesian matching technique on the alternative representations: intensity differences and optical flow. The particular optical flow algorithm used in our experiment was that of Wang & Adelson [14]. For each method, the eigenspace analysis was used to derive corresponding density estimates for the intra/extra classes and recognition proceeded exactly as described in the previous section.

Since it is difficult to compare recognition and false match rates directly (due to the different dimensionalities of $d(I_1, I_2)$ in each case) we systematically varied the dimensions of the principal subspaces M_I and M_E, as in Figure 10 for each method and analyzed the performance in terms of % correct recognition and the number of false matches. Table 11 shows the mean and maximum values computed over the nearly 2,000 different combinations of M_I and M_E for the three different methods: full XYI-warp, intensity differences (I-diff) and optical flow (XY-flow). These results indicate that XYI-warps are in fact the best representation for classification purposes, with intensity differences being second and optical flow being the least effective representation. We believe the reason optical flow is so ineffective is because it has no intensity information encoded in the representation and also since it essentially yields "garbage" for the extrapersonal class (due to the inability of obtaining good correspondences between two different individuals). Notice how the number of false matches, however, is least with optical flow, possibly because it is quite easy to discriminate between the (essentially "garbage") flow field of an extrapersonal warp and that of an intrapersonal one. Also note that in terms of false matches, intensity differences seem to yield worse results than XYI-warps.

5 Conclusions

We have proposed an alternative technique for direct visual matching of images for purposes of recognition and database search. 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 (*e.g.,* template matching) or subspace-restricted norms (*e.g.,* eigenspace matching). This probabilistic framework is also 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 $P(\Omega_I|d(I_1, I_2)$, we obtain an optimal non-linear decision rule for matching and recognition. This aspect of our approach differs from methods which use linear discriminant analysis techniques for visual object recognition (*e.g.,* [15]).

Furthermore, we have experimentally shown that our deformable XYI warping method for obtaining pixel correspondences does indeed lead to an effective representation for $d(I_1, I_2)$, especially when compared with simpler methods such as intensity differences and optical flow. In fact, these methods can essentially be viewed as limiting cases of our general XYI warping method and therefore lack full correspondence: the intensity difference method requires pre-established spatial correspondence between I_1 and I_2, whereas optical flow assumes that I_1 and I_2 only differ by an XY deformation. The XYI warping method, on the other hand, makes no such assumptions and efficiently solves for both types of correspondences in a unified framework. The resultant modal amplitude spectra of these deformations will therefore encode both shape (spatial) and texture (intensity) variations between the two images. The experimental results indicate that a $d(I_1, I_2)$ representation based on full XYI correspondence (*i.e.,* precise alignment/correspondence) does in fact lead to the best overall recognition performance.

References

[1] K. J. Bathe. *Finite Element Procedures in Engineering Analysis.* Prentice-Hall, 1982.

[2] David Beymer. Vectorizing face images by interleaving shape and texture computations. A.I. Memo No. 1537,

Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1995.

[3] I. Craw and P. Cameron. Face recognition by computer. In D. Hogg and R. Boyle, editors, *Proc. British Machine Vision Conference*, pages 498–507. Springer-Verlag, 1992.

[4] I. Craw and et al. Automatic face recognition: Combining configuration and texture. In Martin Bichsel, editor, *Proc. Int'l Workshop on Automatic Face and Gesture Recognition*, Zurich, 1995.

[5] B.K.P. Horn and G. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

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

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

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

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

[10] C. Nastar. Vibration modes for nonrigid motion analysis in 3D images. In *Proceedings of the Third European Conference on Computer Vision (ECCV '94)*, Stockholm, May 1994.

[11] C. Nastar and N. Ayache. Fast segmentation, tracking, and analysis of deformable objects. In *IEEE Proceedings of the Third International Conference on Computer Vision (ICCV'93)*, Berlin, May 1993.

[12] C. Nastar and A. Pentland. Matching and recognition using deformable intensity surfaces. In *IEEE International Symposium on Computer Vision*, Coral Gables, USA, November 1995.

[13] A. Pentland and S. Sclaroff. Closed-form solutions for physically based shape modelling and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-13(7):715–729, July 1991.

[14] J.Y.A. Wang and E.H. Adelson. Representing moving images with layers. *IEEE Transactions on Image Processing*, 3(5):625–638, September 1994.

[15] J.J. Weng. On comprehensive visual learning. In *Proc. NSF/ARPA Workshop on Performance vs. Methodology in Computer Vision*, Seattle, WA, June 1994.