M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 326
Appears in: *The 5th International Conference on Computer Vision, Cambridge, MA, June 1995.*

Probabilistic Visual Learning for Object Detection

Baback Moghaddam and Alex Pentland

Vision and Modeling Group, The Media Laboratory
Massachusetts Institute of Technology
20 Ames St., Cambridge, MA 02139

Abstract

We present an unsupervised technique for visual learning which is based on density estimation in high-dimensional spaces using an eigenspace decomposition. Two types of density estimates are derived for modeling the training data: a multivariate Gaussian (for a unimodal distribution) and a multivariate Mixture-of-Gaussians model (for multimodal distributions). These probability densities are then used to formulate a maximum-likelihood estimation framework for visual search and target detection for automatic object recognition. This learning technique is tested in experiments with modeling and subsequent detection of human faces and non-rigid objects such as hands.

1 Introduction

The standard detection paradigm in image processing is that of normalized correlation or template matching. However this approach is only optimal in the simplistic case of a *deterministic* signal embedded in additive white Gaussian noise. When we begin to consider a target *class* detection problem — *e.g.,* finding a generic human face in a scene — we must incorporate the underlying probability distribution of the object. Subspace methods and eigenspace decompositions are particularly well-suited to such a task since they provide a compact and parametric description of the object's appearance and also automatically identify the *degrees-of-freedom* of the underlying statistical variability.

In particular, the eigenspace formulation leads to a powerful alternative to standard detection techniques such as template matching or normalized correlation. The reconstruction error (or residual) of the eigenspace decomposition (referred to as the "distance-from-face-space" in the context of the work with "eigenfaces" [14]) is an effective indicator of similarity. The residual error is easily computed using the projection coefficients and the original signal energy. This detection strategy is equivalent to matching with a linear combination of *eigentemplates* and allows for a greater range of distortions in the input signal (including lighting, and moderate rotation and scale). In a statistical signal detection framework, the use of eigentemplates has been shown to yield superior performance in comparison with standard matched filtering [6][10].

Pentland *et al.* [10] used this formulation for a modular eigenspace representation of facial features where the corresponding residual — referred to as "distance-from-*feature*-space" (DFFS) — was used for localization and detection. Given an input image, a saliency map was constructed by computing the DFFS at each pixel. When using M eigenvectors, this requires M convolutions (which can be efficiently computed using an FFT) plus an additional local energy computation. The global minimum of this distance map was then selected as the best estimate of the target location.

We will show that the DFFS can be interpreted as an estimate of a marginal component of the probability density of the object in image space and that a complete estimate must also incorporate a second marginal density based on a complementary "distance-*in*-feature-space" (DIFS). Using the probability density of the object, we formulate the problem of target detection in a maximum likelihood (ML) estimation framework.

2 Density Estimation

Our approach to automatic visual learning is based on density estimation. However, instead of applying estimation techniques directly to the original high-dimensional space of the imagery, we use an eigenspace decomposition to yield a computationally feasible estimate. Specifically, given a set of training images $\{\mathbf{x}^t\}_{t=1}^{N_T}$, from an object class Ω, we wish to estimate the class membership or *likelihood* function for this data — *i.e.,* $P(\mathbf{x}|\Omega)$. In this section, we examine two density estimation techniques for visual learning of high-dimensional data. The first method is based on the assumption of a Gaussian distribution while the second method generalizes to arbitrarily complex distributions using a Mixture-of-Gaussians density model. Before introducing these estimators we briefly review eigenvector decomposition as commonly used in principal component analysis (PCA) [5].

2.1 Principal Component Imagery

Given a set of m-by-n images $\{I^t\}_{t=1}^{N_T}$, we can form a training set of vectors $\{\mathbf{x}^t\}$, where $\mathbf{x} \in \mathcal{R}^{N=mn}$, by lexicographic ordering of the pixel elements of each image I^t. The basis functions in a Karhunen-Loeve Transform (KLT) [7] are obtained by solving the eigenvalue problem

$$\Lambda = \Phi^T \Sigma \Phi \tag{1}$$

where Σ is the covariance matrix of the data, Φ is the eigenvector matrix of Σ and Λ is the corresponding diagonal matrix of eigenvalues. In PCA, a partial KLT is performed to identify the largest-eigenvalue eigenvectors and obtain a principal component feature vector $\mathbf{y} = \Phi_{\mathrm{M}}^T \tilde{\mathbf{x}}$, where

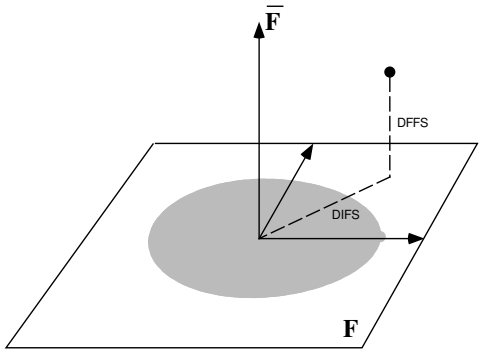


Figure 1: The principal subspace F and its orthogonal complement $\bar{F}$ for a Gaussian density.

$\tilde{\mathbf{x}} = \mathbf{x} - \bar{\mathbf{x}}$ is the mean-normalized image vector and Φ_M is a submatrix of Φ containing the principal eigenvectors. PCA can be seen as a linear transformation $\mathbf{y} = \mathcal{T}(\mathbf{x}) : \mathcal{R}^N \rightarrow \mathcal{R}^M$ which extracts a lower-dimensional subspace of the KL basis corresponding to the maximal eigenvalues. This corresponds to an orthogonal decomposition of the vector space $\mathcal{R}^N$ into two mutually exclusive and complementary subspaces: the principal subspace (or feature space) $F = \{\Phi_i\}_{i=1}^{M}$ containing the principal components and its orthogonal complement $\bar{F} = \{\Phi_i\}_{i=M+1}^{N}$, as illustrated in Figure 1.

In a partial KL expansion, the residual reconstruction error is defined as

$$\epsilon^2(\mathbf{x}) \;=\; \sum_{i=M+1}^{N} y_i^2 \;=\; ||\tilde{\mathbf{x}}||^2 \;-\; \sum_{i=1}^{M} y_i^2 \qquad (2)$$

and can be easily computed from the first M principal components and the L_2-norm of the mean-normalized image $\tilde{\mathbf{x}}$. Consequently the L_2 norm of every element $\mathbf{x} \in \mathcal{R}^N$ can be decomposed in terms of its projections in these two subspaces. We refer to the component in the orthogonal subspace $\bar{F}$ as the "distance-from-feature-space" (DFFS) which is a simple Euclidean distance and is equivalent to the residual error $\epsilon^2(\mathbf{x})$ in Eq.(2). The component of $\mathbf{x}$ which lies *in* the feature space F is referred to as the "distance-in-feature-space" (DIFS) but is generally not a distance-based norm, but can be interpreted in terms of the probability distribution of y in F.

2.2 Gaussian F-Space Densities

We begin by considering an optimal approach for estimating high-dimensional Gaussian densities. We assume that we have (robustly) estimated the mean $\bar{\mathbf{x}}$ and covariance Σ of the distribution from the given training set $\{\mathbf{x}^t\}$. Under this assumption, the likelihood of a input pattern $\mathbf{x}$ is given by

$$P(\mathbf{x}|\Omega) = \frac{\exp\left[-\frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}})^T\Sigma^{-1}(\mathbf{x}-\bar{\mathbf{x}})\right]}{(2\pi)^{N/2}\;|\Sigma|^{1/2}} \qquad (3)$$

The sufficient statistic for characterizing this likelihood is the *Mahalanobis* distance

$$d(\mathbf{x}) \;=\; \tilde{\mathbf{x}}^T\Sigma^{-1}\tilde{\mathbf{x}} \qquad (4)$$

where $\tilde{\mathbf{x}} = \mathbf{x} - \bar{\mathbf{x}}$. Using the eigenvectors and eigenvalues of Σ we can rewrite Σ^{-1} in the diagonalized form

$$\begin{aligned} d(\mathbf{x}) &= \tilde{\mathbf{x}}^T\Sigma^{-1}\tilde{\mathbf{x}} \\ &= \tilde{\mathbf{x}}^T\left[\Phi\Lambda^{-1}\Phi^T\right]\tilde{\mathbf{x}} \\ &= \mathbf{y}^T\Lambda^{-1}\mathbf{y} \end{aligned} \qquad (5)$$

where $\mathbf{y} = \Phi^T\tilde{\mathbf{x}}$ are the new variables obtained by the change of coordinates in a KLT. Because of the diagonalized form, the *Mahalanobis* distance can also be expressed in terms of the sum

$$d(\mathbf{x}) \;=\; \sum_{i=1}^{N}\frac{y_i^2}{\lambda_i} \qquad (6)$$

We now seek to estimate $d(\mathbf{x})$ using only the M principal projections. Therefore, we formulate an estimator for $d(\mathbf{x})$ as follows

$$\begin{aligned} \hat{d}(\mathbf{x}) &= \sum_{i=1}^{M}\frac{y_i^2}{\lambda_i} \;+\; \frac{1}{\rho}\left[\sum_{i=M+1}^{N} y_i^2\right] \\ &= \sum_{i=1}^{M}\frac{y_i^2}{\lambda_i} \;+\; \frac{1}{\rho}\,\epsilon^2(\mathbf{x}) \end{aligned} \qquad (7)$$

where the term in the brackets is the DFFS $\epsilon^2(\mathbf{x})$, which as we have seen can be computed using the first M principal components. We can therefore write the form of the likelihood estimate based on $\hat{d}(\mathbf{x})$ as the product of two marginal and independent Gaussian densities

$$\begin{aligned} \hat{P}(\mathbf{x}|\Omega) &= \left[\frac{\exp\left(-\frac{1}{2}\sum_{i=1}^{M}\frac{y_i^2}{\lambda_i}\right)}{(2\pi)^{M/2}\prod_{i=1}^{M}\lambda_i^{1/2}}\right]\cdot\left[\frac{\exp\left(-\frac{\epsilon^2(\mathbf{x})}{2\rho}\right)}{(2\pi\rho)^{(N-M)/2}}\right] \\ &= P_F(\mathbf{x}|\Omega)\;\hat{P}_{\bar{F}}(\mathbf{x}|\Omega) \end{aligned} \qquad (8)$$

where $P_F(\mathbf{x}|\Omega)$ is the true marginal density in F-space and $\hat{P}_{\bar{F}}(\mathbf{x}|\Omega)$ is the estimated marginal density in the orthogonal complement $\bar{F}$-space. The optimal value of ρ can now be determined by minimizing a suitable cost function $J(\rho)$. From an information-theoretic point of view, this cost function should be the Kullback-Leibler divergence [3] between the true density $P(\mathbf{x}|\Omega)$ and its estimate $\hat{P}(\mathbf{x}|\Omega)$

$$J(\rho) \;=\; \mathrm{E}\left[\log\frac{P(\mathbf{x}|\Omega)}{\hat{P}(\mathbf{x}|\Omega)}\right] \qquad (9)$$

Using the diagonalized forms of the *Mahalanobis* distance $d(\mathbf{x})$ and its estimate $\hat{d}(\mathbf{x})$ and the fact that $\mathrm{E}[y_i^2] = \lambda_i$, it can be easily shown that

$$J(\rho) \;=\; \frac{1}{2}\sum_{i=M+1}^{N}\left[\frac{\lambda_i}{\rho} - 1 + \log\frac{\rho}{\lambda_i}\right] \qquad (10)$$

The optimal weight ρ^* can be then found by minimizing this cost function with respect to ρ. Solving the equation $\frac{\partial J}{\partial \rho} = 0$ yields

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

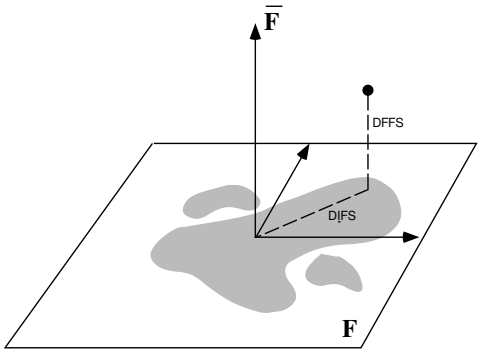


Figure 2: The principal subspace F and its orthogonal complement $\bar{F}$ for an arbitrary density.

which is simply the arithmetic average of the eigenvalues in the orthogonal subspace $\bar{F}$. In addition to its optimality, ρ^* also results in an *unbiased* estimate of the *Mahalanobis* distance — *i.e,* $\mathrm{E}[\hat{d}(\mathbf{x};\rho^*)] = \mathrm{E}[d(\mathbf{x})]$. What this derivation shows is that once we select the M-dimensional principal subspace F (as indicated, for example, by PCA), the optimal density estimate $\hat{P}(\mathbf{x}|\Omega)$ has the form of Eq.(8) with ρ given by Eq.(11).

2.3 Multimodal F-space Densities

When the training set represents multiple views or multiple objects under varying illumination conditions, the distribution of training views in F-space is no longer unimodal. In fact the training data tends to lie on complex and non-separable low-dimensional manifolds in image space [1]. One way to tackle this multimodality is to build a view-based (or object-based) formulation where separate eigenspaces are used for each view [10]. Another approach is to capture the complexity of these manifolds in a universal or *parametric* eigenspace using splines [9], or local basis functions [2].

If we assume that the $\bar{F}$-space components are Gaussian and independent of the principal features in F (this would be true in the case of pure observation noise in $\bar{F}$) we can still use the separable form of the density estimate $\hat{P}(\mathbf{x}|\Omega)$ in Eq.(8) where $P_F(\mathbf{x}|\Omega)$ is now an *arbitrary* density $P(\mathbf{y})$ in the principal component vector $\mathbf{y}$. Figure 2 illustrates the decomposition, where the DFFS is the residual $\epsilon^2(\mathbf{x})$ as before. The DIFS, however, is no longer a simple *Mahalanobis* distance but can nevertheless be interpreted as a "distance" by relating it to $P(\mathbf{y})$ — *e.g.,* as DIFS $= -\log P(\mathbf{y})$.

The density $P(\mathbf{y})$ can be estimated using a parametric mixture model. Specifically, we can model arbitrarily complex densities using a Mixture-of-Gaussians

$$P(\mathbf{y}|\Theta) \;=\; \sum_{i=1}^{N_c} \pi_i \; g(\mathbf{y};\mu_i,\Sigma_i) \qquad (12)$$

where $g(\mathbf{y};\mu,\Sigma)$ is an M-dimensional Gaussian density with mean vector μ and covariance Σ, and the π_i are the mixing parameters of the components, satisfying $\sum \pi_i = 1$. The mixture is completely specified by the parameter $\Theta = \{\pi_i,\mu_i,\Sigma_i\}_{i=1}^{N_c}$. Given a training set $\{\mathbf{y}^t\}_{t=1}^{N_T}$ the mixture parameters can be estimated using the ML principle

$$\Theta^* \;=\; \mathrm{argmax}\left[\prod_{t=1}^{N_T} P(\mathbf{y}^t|\Theta)\right] \qquad (13)$$

This estimation problem is best solved using the Expectation-Maximization (EM) algorithm [4]. The EM algorithm is monotonically convergent in *likelihood* and is thus guaranteed to find a local maximum in the total likelihood of the training set. Further details of the EM algorithm for estimation of mixture densities can be found in [12].

Given our operating assumptions — that the training data is truly M-dimensional (at most) and resides solely in the principal subspace F with the exception of perturbations due to white Gaussian measurement noise, or equivalently that the $\bar{F}$-space component of the data is itself a separable Gaussian density — the estimate of the complete likelihood function $P(\mathbf{x}|\Omega)$ is given by

$$\hat{P}(\mathbf{x}|\Omega) \;=\; P(\mathbf{y}|\Theta^*)\;\hat{P}_{\bar{F}}(\mathbf{x}|\Omega) \qquad (14)$$

where $\hat{P}_{\bar{F}}(\mathbf{x}|\Omega)$ is a Gaussian component density based on the DFFS, as before.

3 Maximum Likelihood Detection

The density estimate $\hat{P}(\mathbf{x}|\Omega)$ can be used to compute a local measure of target saliency at each spatial position (i,j) in an input image based on the vector $\mathbf{x}$ obtained by the lexicographic ordering of the pixel values in a local neighborhood R_{ij} — *i.e.,* $S(i,j;\Omega) = \hat{P}(\mathbf{x}|\Omega)$ where $\mathbf{x}$ is the vectorized region R_{ij}. The ML estimate of position of the target Ω is then given by

$$(i,j)^{\mathrm{ML}} \;=\; \mathrm{argmax}\; S(i,j;\Omega) \qquad (15)$$

Similarly, we can extend the parameter space to include scale, resulting in *multiscale* saliency maps. The likelihood computation is performed (in parallel) on linearly scaled versions of the input image $I^{(\sigma)}$ corresponding to a predetermined set of (linearly spaced) scales $\{\sigma_1,\sigma_2,\cdots\sigma_n\}$

$$S(i,j,k;\;\Omega) = \hat{P}(\mathbf{x}^{ijk}|\Omega) \qquad (16)$$

where $\mathbf{x}^{ijk}$ is the vector obtained from a local subimage in the multiscale representation. The ML estimate of the spatial position and scale of the object is then defined as

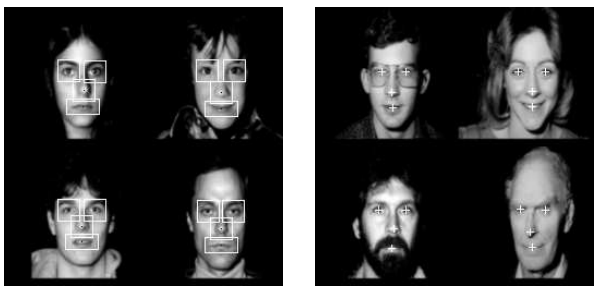
$$(i,j,k)^{\mathrm{ML}} \;=\; \mathrm{argmax}\; S(i,j,k;\;\Omega) \qquad (17)$$

4 Applications

The above ML detection technique has been tested in the detection of complex natural objects including human faces, facial features (*e.g.,* eyes), as well as non-rigid and articulated objects such as human hands. In this section we will present several examples from these application domains.

4.1 Faces

The eigentemplate approach to the detection of facial features in "mugshots" was proposed in [10], where the DFFS metric was shown to be superior to standard template



(a) (b)

Figure 3: (a) Examples of facial feature training templates and (b) the resulting typical detections.

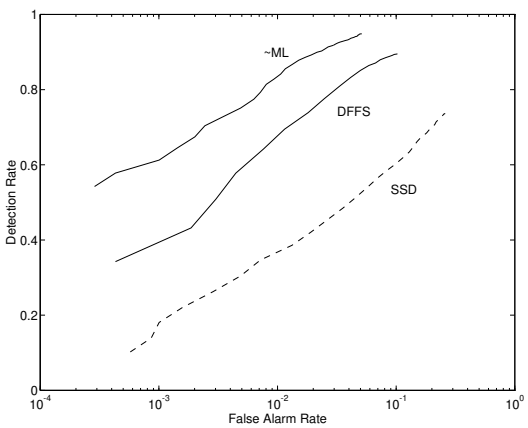


Figure 4: Performance of an SSD, DFFS and a ML detector.

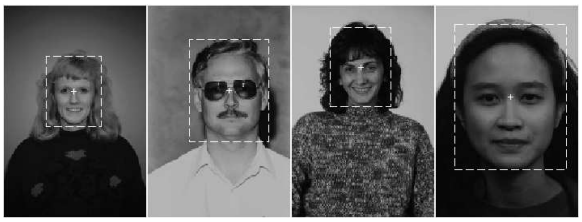
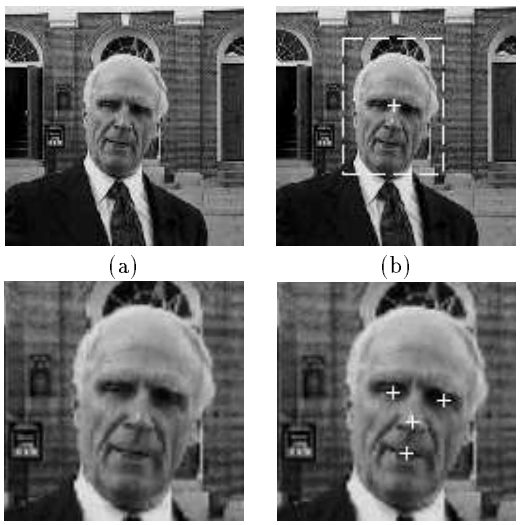


Figure 5: Examples of multiscale face detection.



(a) (b) (c) (d)

Figure 7: (a) original image, (b) position and scale estimate, (c) normalized head image, (d) position of facial features.

matching for target detection. The detection task was the estimation of the position of facial features (the left and right eyes, the tip of the nose and the center of the mouth) in frontal view photographs of faces at fixed scale. Figure 3 shows examples of facial feature training templates and the resulting detections on the MIT Media Laboratory's database of 7,562 "mugshots".

We have compared the detection performance of three different detectors on approximately 7,000 test images from this database: a sum-of-square-differences (SSD) detector based on the average facial feature (in this case the left eye), an eigentemplate or DFFS detector and a ML detector based on $S(i,j;\Omega)$ as defined in section 3 and using a unimodal F-space density as in section 2.2. Figure 4(a) shows the *receiver operating characteristic* (ROC) curves for these detectors, obtained by varying the detection threshold independently for each detector. The DFFS and ML detectors were computed based on a 5-dimensional principal subspace. Since the projection coefficients were unimodal a Gaussian distribution was used in modeling the true distribution for the ML detector as in section 2.2. Note that the ML detector exhibits the best detection vs. false-alarm tradeoff and yields the highest detection rate (of 95%). Indeed, at the *same* detection rate the ML detector has a false-alarm rate which is nearly 2 orders of magnitude lower than the SSD.

We have also incorporated and tested the multiscale version of the ML detection technique in a face detection task. This multiscale head finder was tested on the ARPA FERET database where 97% of 2,000 faces were correctly detected. Figure 5 shows examples of the ML estimate of the position and scale on these images. The multiscale saliency maps $S(i,j,k;\Omega)$ were computed based on the likelihood estimate $\hat{P}(\mathbf{x}|\Omega)$ in a 10-dimensional principal subspace using a Gaussian model (section 2.2). Note that this detector is able to localize the position and scale of the head despite variations in hair style and hair color, as well as presence of sunglasses. Illumination invariance was obtained by normalizing the input subimage $\mathbf{x}$ to a zero-mean unit-norm vector.

This multiscale face detector has also been used as the *attentional* component of an automatic system for recognition and model-based coding of faces. The block diagram of this system is shown in Figure 6 which consists of a two-stage object detection and alignment stage, a contrast normalization stage, and a feature extraction stage whose output is used for both recognition and coding.

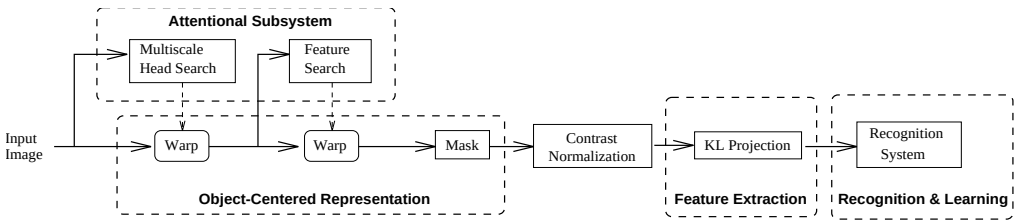


Figure 6: The face processing system.

(a) (b) (c)

Figure 8: (a) aligned face, (b) eigenspace reconstruction (85 bytes) (c) JPEG reconstruction (530 bytes).

Figure 7 illustrates the operation of the detection and alignment stage on a natural test image containing a human face.

The first step in this process is illustrated in Figure 7(b) where the ML estimate of the position and scale of the face are indicated by the cross-hairs and bounding box. Once these regions have been identified, the estimated scale and position are used to normalize for translation and scale, yielding a standard "head-in-the-box" format image (Figure 7(c)). A second feature detection stage operates at this fixed scale to estimate the position of 4 facial features: the left and right eyes, the tip of the nose and the center of the mouth (Figure 7(d)). Once the facial features have been detected, the face image is warped to align the geometry and shape of the face with that of a canonical model. Then the facial region is extracted (by applying a fixed mask) and subsequently normalized for contrast. The geometrically aligned and normalized image (shown in Figure 8(a)) is then projected onto a custom set of eigenfaces to obtain a feature vector which is then used for recognition purposes as well as facial image coding.

Figure 8 shows the normalized facial image extracted from Figure 7(d), its reconstruction using a 100-dimensional eigenspace representation (requiring only 85 bytes to encode) and a comparable non-parametric reconstruction obtained using a standard transform-coding approach for image compression (requiring 530 bytes to encode). This example illustrates that the eigenface representation used for recognition is also an effective *model-based* representation for data compression. The first 8 eigenfaces used for this representation are shown in Figure 9.

Figure 10 shows the results of a similarity search in an image database tool called Photobook [11]. Each face

Figure 9: The first 8 eigenfaces.

Figure 10: Photobook: FERET face database.

in the database was automatically detected and aligned by the face processing system in Figure 6. The normalized faces were then projected onto a 100-dimensional eigenspace. The image in the upper left is the one searched on and the remainder are the ranked nearest neighbors in the FERET database. The top three matches in this case are images of the same person taken a month apart and at different scales. The recognition accuracy (defined as the percent correct rank-one matches) on a database of 155 individuals is 99% [8].

We have also extended the normalized eigenface representation into an edge-based domain for facial description. We simply run the normalized facial image through a Canny edge detector to yield an edge-map. Unfortunately binary edge maps, are highly uncorrelated with each other due to their sparse nature, and therefore result in a very high-dimensional principal subspace. Therefore, to reduce the intrinsic dimensionality, we *induced* spatial correlation via a *diffusion* process on the binary edge map, which effectively broadens and "smears" the edges, yielding a continuous-valued edge-map as shown in Figure 11(a).

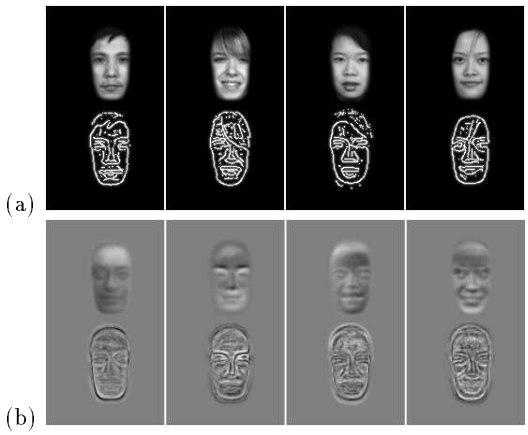


Figure 11: (a) Examples of combined texture/edge-based face representations and (b) few of the resulting eigenvectors.

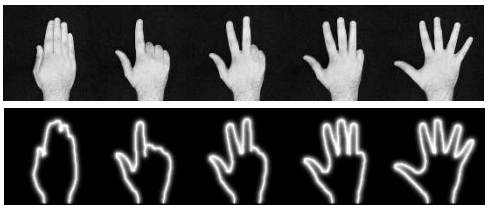


Figure 12: (a) Examples of hand gestures and (b) their diffused edge-based representation.

Such an edge-map is simply an alternative representation which imparts mostly *shape* (as opposed to texture) information and has the distinct advantage of being less susceptible to illumination changes. The recognition rate of a pure edge-based normalized eigenface representation (on the same database of 155 individuals) was found to be 95% which is surprising considering that it utilizes what appears to be (to humans at least) a rather impoverished representation. The slight drop in recognition rate is most likely due to the increased dimensionality of this representation space and its greater sensitivity to expression changes, *etc.*

Interestingly, we can combine both texture and edge-based representations of the object by simply performing a KL expansion on the augmented images shown in Figure 11. The resulting principal eigenvectors conveniently decorrelate the joint representation and provide a basis set which optimally spans both domains simultaneously. With this bimodal representation, the recognition rate was found to be 97%. Though still less than a normalized grayscale representation, we believe a bimodal representation can have distinct advantages for tasks other than recognition, such as detection and image interpolation.

4.2 Hands

We have also applied our eigenspace density estimation technique to articulated and non-rigid objects such as hands. In this particular domain, however, the normalized

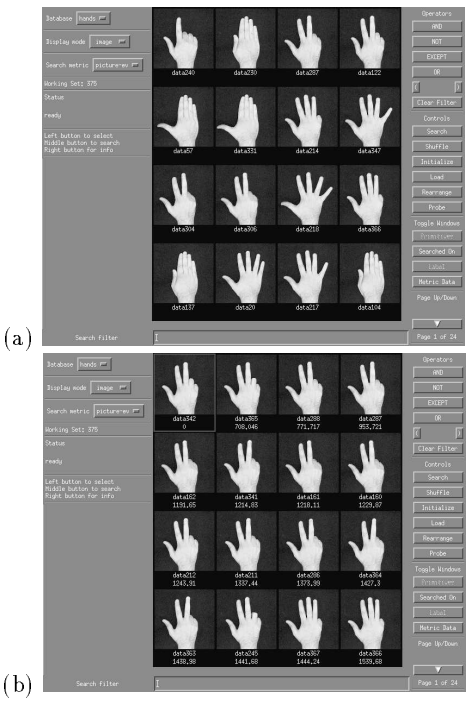


Figure 13: (a) A random assortment of hand gestures (b) images ordered by similarity (left-to-right, top-to-bottom) to the image at the upper left.

grayscale image is an unsuitable representation since, unlike faces, hands are essentially textureless objects. Their identity is characterized by the variety of *shapes* they can assume. For this reason we have chosen an edge-based representation of hand shapes which is invariant to illumination, contrast and scene background. A training video sequence of hand gestures was obtained against a black background. The 2D contour of the hand was then extracted using a Canny edge-operator and diffused as in the case of facial edge maps (see Figure 12). We note that this *spatiotopic* representation of shape is biologically motivated and is different from shape representations which are based on computational considerations (*e.g.,* Fourier descriptors and "snakes").

It is important to verify whether such a representation is valid for modeling hand shapes. Therefore we tested the diffused contour image representation in a recognition experiment which yielded a 100% rank-one accuracy on the 375-frame image sequence containing multiple examples of 7 hand gestures. The matching technique was a nearest-neighbor classification rule in a 16-dimensional principal subspace. Figure 13(a) shows some examples of the various hand gestures used in this experiment. Figure 13(b) shows the 15 images that are most similar to the "two" gesture appearing in the top left. Note that the hand gestures judged most similar are all objectively the same gesture.

Naturally, the success of such a recognition system is critically dependent on the ability to find the hand (in any of its articulated states) in a cluttered scene, to account for its scale and to align it with respect to an object-centered reference frame prior to recognition. This localization

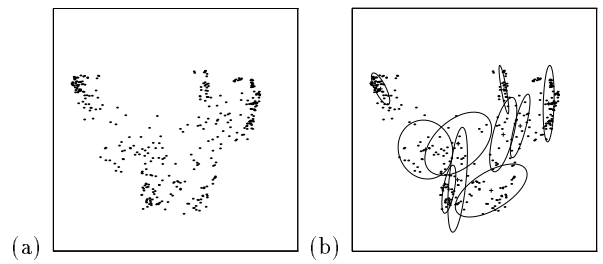


Figure 14: (a) Distribution of training hand shapes (shown in the first two dimensions of the principal subspace) (b) Mixture-of-Gaussians representation using 10 components.

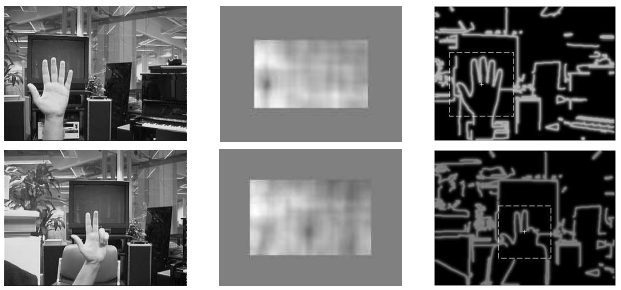


Figure 15: (a) Original grayscale image, (b) negative log-likelihood map (at most likely scale) and (c) ML estimate of position and scale superimposed on edge-map.

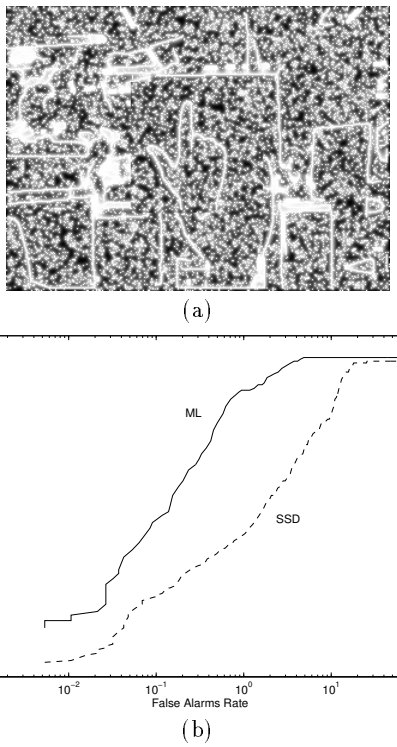


Figure 16: (a) Example of test frame containing a hand gesture amidst severe background clutter and (b) ROC curve performance contrasting SSD and ML detectors.

can be achieved with the same multiscale ML detection paradigm used with faces, with the exception that the underlying image representation of the hands is a diffused edge map rather than the original grayscale image.

The probability distribution of hand shapes in this representation was automatically learned using our eigenspace density estimation technique. In this case, however, the distribution of training data is *multimodal* due to the different hand shapes for each gesture. Therefore the multimodal density estimation technique in section 2.3 was used. Figure 14(a) shows a projection of the training data on the first two dimensions of the principal subspace F (defined in this case by $M = 16$) which exhibit the underlying multimodality of the data. Figure 14(b) shows a 10-component Mixture-of-Gaussians density estimate for the training data. The parameters of this estimate were obtained with 20 iterations of the EM algorithm. The orthogonal $\bar{F}$-space component of the density was modeled with a Gaussian distribution as in section 2.3.

The resulting complete density estimate $\hat{P}(\mathbf{x}|\Omega)$ was then used in a detection experiment on test imagery of hand gestures against a cluttered background scene. In accordance with our representation, the input imagery was first pre-processed to generate a diffused edge map and then scaled accordingly for a multiscale saliency computation. Figure 15 shows two examples from the test sequence, where we have shown the original image, the negative log-likelihood saliency map, and the ML estimates of position and scale. Note that these examples represent two different hand gestures at slightly different scales.

To quantify the performance of the ML detector on hands we carried out the following experiment. The original 375-frame video sequence of training hand gestures was divided into 2 parts. The first (training) half of this sequence was used for learning, including computation of the KL basis and the subsequent EM clustering. For this experiment we used a 5-component mixture in a 10-dimensional principal subspace. The 2nd (testing) half of the sequence was then embedded in the background scene, which contains a variety of shapes. In addition, severe noise conditions were simulated as shown in Figure 16(a).

We then compared the detection performance of an SSD detector (based on the mean edge-based hand representation) and a probabilistic detector based on the complete estimated density. The resulting negative-log-likelihood detection maps were passed through a valley-detector to isolate local minimum candidates which were then subjected to a ROC analysis. A correct detection was defined as a below-threshold local minimum within a 5-pixel radius of the ground truth target location. Figure 16(b) shows the performance curves obtained for the two detectors. We note, for example, that at an 85% detection probability the ML detector yields (on the average) 1 false alarm per frame, where as the SSD detector yields an order of magnitude more false alarms.

5 Discussion

We have described a density estimation technique for unsupervised visual learning which exploits the *intrinsic* low-dimensionality of the training imagery to form a computationally simple estimator for the complete likelihood function of the object. Our estimator is based on a subspace decomposition and can be evaluated using only the M-dimensional principal component vector. We have derived the form for an optimal estimator and its associated expected cost for the case of a Gaussian density. In contrast to previous work on learning and characterization — which uses PCA primarily for dimensionality reduction and/or feature extraction — our method uses the eigenspace decomposition as an integral part of estimating *complete* density functions in high-dimensional image spaces. These density estimates were then used in a maximum likelihood formulation for target detection. The multiscale version of this detection strategy was demonstrated in applications in which it functioned as an attentional subsystem for object recognition. The performance was found to be superior to existing detection techniques in experimental results on a large number of test data (on the order of thousands).

We conclude by noting that from a probabilistic perspective, the class-conditional density $P(\mathbf{x}|\Omega)$ is the most important data representation to be learned. This density is the critical component in detection, recognition, prediction, interpolation and general inference. For example, having learned these densities for several object classes $\{\Omega_1, \Omega_2, \cdots, \Omega_n\}$, one can invoke a Bayesian framework for classification and recognition:

$$P(\Omega_i|\mathbf{x}) = \frac{P(\mathbf{x}|\Omega_i)P(\Omega_i)}{\sum_{j=1}^{n} P(\mathbf{x}|\Omega_j)P(\Omega_j)} \qquad (18)$$

Such a framework is also important in detection. In fact, the ML detection framework can be extended using the notion of a "not-class" $\bar{\Omega}$, resulting in *a posteriori* saliency maps of the form

$$P(\Omega|\mathbf{x}) = \frac{P(\mathbf{x}|\Omega)P(\Omega)}{P(\mathbf{x}|\bar{\Omega})P(\bar{\Omega}) + P(\mathbf{x}|\Omega)P(\Omega)} \qquad (19)$$

where now a maximum *a posteriori* (MAP) rule can be used to estimate the position and scale of the object. One difficulty with such a formulation is that the "not-class" $\bar{\Omega}$ is, in practice, too broad a category and is therefore multimodal and very high-dimensional. One possible approach to this problem is to use ML detection to identify the particular subclass of $\bar{\Omega}$ which has high likelihoods (*e.g.,* typical false alarms) and then to estimate this distribution and use it in the MAP framework. This can be viewed as a probabilistic approach to learning using positive as well as *negative* examples. The use of negative examples has been shown to be critically important in building robust face detection systems [13].

Acknowledgements

The FERET face database was provided by the US Army Research Laboratory. This research was partially funded by British Telecom.

References

[1] Bichsel, M., and Pentland, A., "Human Face Recognition and the Face Image Set's Topology," *CVGIP: Image Understanding*, Vol. 59, No. 2, pp. 254-261, 1994.

[2] Bregler, C., and Omohundro, S.M., "Surface learning with applications to lip reading," in *Advances in Neural Information Processing Systems 6*, eds. J.D. Cowan, G. Tesauro and J. Alspector, Morgan Kaufman Publishers, San Fransisco, pp. 43-50, 1994.

[3] Cover, M. and Thomas, J.A., *Elements of Information Theory*, John Wiley & Sons, New York, 1994.

[4] Dempster, A.P., Laird, N.M., Rubin, D.B., "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society B,* vol. 39, 1977.

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

[6] Kumar, B., Casasent, D., and Murakami, H., "Principal Component Imagery for Statistical Pattern Recognition Correlators," *Optical Engineering,* vol. 21, no. 1, Jan/Feb 1982.

[7] Loeve, M.M., *Probability Theory*, Van Nostrand, Princeton, 1955.

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

[9] Murase, H., and Nayar, S. K., "Learning and Recognition of 3D Objects from Appearance" in *IEEE 2nd Qualitative Vision Workshop*, New York, NY, June 1993.

[10] Pentland, A., Moghaddam, B. and Starner, T., "View-based and modular eigenspaces for face recognition," *Proc. of IEEE Conf. on Computer Vision & Pattern Recognition*, June 1994.

[11] Pentland, A., Picard, R., and Sclaroff, S., "Photobook: Tools for Content-Based Manipulation of Image Databases," in *Storage and Retrieval of Image and Video Databases II*, SPIE vol. 2185, 1994.

[12] Redner, R.A., and Walker, H.F., "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Review*, vol. 26, no. 2, pp. 195-239, 1984.

[13] Sung, K., and Poggio, T., "Example-based Learning for View-based Human Face Detection," A.I. Memo No. 1521, Artificial Intelligence Laboratory, MIT, 1994.

[14] Turk, M., and Pentland, A., "Eigenfaces for Recognition," *Journal of Cognitive Neuroscience,* Vol. 3, No. 1, 1991.