Appeared in: Proceedings of European Conference on Computer Vision, 2002.

Face recognition from long-term observations

Gregory Shakhnarovich, John W. Fisher, and Trevor Darrell

Artificial Intelligence Laboratory
Massachusetts Institute of Technology
{gregory,fisher,trevor}@ai.mit.edu

Abstract. We address the problem of face recognition from a large set of images obtained over time - a task arising in many surveillance and authentication applications. A set or a sequence of images provides information about the variability in the appearance of the face which can be used for more robust recognition. We discuss different approaches to the use of this information, and show that when cast as a statistical hypothesis testing problem, the classification task leads naturally to an information-theoretic algorithm that classifies sets of images using the relative entropy (Kullback-Leibler divergence) between the estimated density of the input set and that of stored collections of images for each class. We demonstrate the performance of the proposed algorithm on two medium-sized data sets of approximately frontal face images, and describe an application of the method as part of a view-independent recognition system.

1 Introduction

Recognition in the context of visual surveillance applications is a topic of growing interest in computer vision. Face recognition has generally been posed as the problem of recognizing an individual from a single "mug shot", and many successful systems have been developed (e.g. Visionics, Bochum/USC). Separately, systems for tracking people in unconstrained environments have become increasingly robust, and are able to track individuals for minutes or hours [8, 5]. These systems typically can provide images of users at low or medium resolution, possibly from multiple viewpoints, over long periods of time. For optimal recognition performance, the information from all images of a user should be included in the recognition process.

In such long-term recognition problems, sets of observations must be compared to a list of known models. Typically the models themselves are obtained from sets of images, and this process can be thought of as a set matching problem. There are many possible schemes for integrating information from multiple observations, using either "early" or "late" integration approaches. Early integration schemes might consist of selecting the "best" observation from each set using some quality metric, or simply averaging together all observations prior to classification. In the case of late integration, the common statistical approach is to take the product of the likelihoods of each observation. This approach, while

simple and well justified under certain conditions, often fails to account for the presence of outliers in the observation data as well as the fact that the data are expected to be observed with a certain variability (e.g., repeated observations of the most likely face of an individual does not yield a typical set of faces for that individual, if we presume human faces occur with some natural variability.)

We propose an approach to face recognition using sets of images, in which we directly compare models of *probability distributions* of observed and model faces. Variability in the observed data is considered when comparing distribution models, and thus outliers and natural variation can be handled properly. Many approaches to distribution modeling and matching are possible within this general framework. In this paper we develop a method based on a Gaussian model of appearance distribution and a matching criterion using the Kullback-Leibler divergence measure. In our method, observations and models can be compared efficiently using a closed-form expression. In addition to being potentially more accurate, our approach can be more efficient than approaches that require comparing each observation with each model example image since our model distributions are compactly parameterized.

We have evaluated our approach on two different data sets of frontal face images. The first was a collection of people observed at a distance in frontal view in an indoor office environment. The second was computed using a recently proposed scheme for integration across multiple viewpoints [15], in which images from several cameras with arbitrary views of the user's face were combined to generate a virtual frontal view. With both data sets, our distribution matching technique offered equal or improved recognition performance compared to traditional approaches, and showed robustness with respect to the choice of model parameters.

Throughout the paper we will assume independence between any two samples in a set of images, thus ignoring, for example, the dynamics of facial expression. While disregarding a potentially important cue , this will also remove the assumption of consecutive frames, making it easier to recognize a person from sparse observations such as are available, for instance, in surveillance systems, where the subject does not face the camera all the time. Furthermore, the training set may be derived from sparse or unordered observations rather than a sequence.

The remainder of the paper is organized as follows. We start with a discussion of previous work on this problem and of known methods appropriate for classifying sets of images. In Sections 3 and 4 we present the statistical analysis leading to our distribution matching algorithm, described in detail in Section 5. Section 6 contains a report on the empirical comparative study of the discussed methods, and is followed by conclusions and discussion in Section 7.

2 Previous work

The area of recognition from a single face image is very well established. Both local, feature-based and global, template-based methods have been extensively

explored [3]. The state of the art today is defined by a family of recognition schemes related to eigendecomposition of the data directly [11] and through Linear Discriminate Analysis [1], or to local feature analysis of images [13, 17], all of which achieve very high performance. However, face recognition from an image sequence, or more generally from a set of images, has been a subject of relatively few published studies.

The common idea in most of the published work is that recognition performance can be improved by modeling the variability in the observed data. Recently proposed detection and tracking algorithms use the dynamics of the consecutive images of a face, and integrate recognition into the tracking framework. In [6], the variation of the individual faces is modeled in the framework of the Active Appearance Model. In [7], the "temporal signature" of a face is defined, and feed-forward neural network is used in order to classify the sequence. In [2], people are recognized from a sequence of rotating head images, using trajectories in a low-dimensional eigenspace. They assume that each image is associated with a known pose – a situation which is uncommon in practice. In all three papers, the temporal constraints of the sequence (i.e., the images being consecutive in time) are crucial. We are interested in recognition under looser assumptions, when the images are not necessarily finely sampled, or even ordered in time.

2.1 Late integration strategies

A number of methods are applicable to the situation where multiple observations of the same person's face are accumulated over time. This is an instance of the more general problem of fusion of evidence from multiple measurements. Kittler *et al*in [9] present a statistical interpretation of a number of common methods for cross- modal fusion, such as the product, maximum , and majority rules, which are also appropriate for late integration over a set of observations from a single modality. For example, it can be shown that under the assumption that the samples in the set are distributed i.i.d., the *product rule* is the *maximum likelihood* classifier of the set:

$$w^* = \underset{w_i}{\operatorname{argmax}} \prod_{t=1}^{n} p(x_t|w_i) = \underset{w_i}{\operatorname{argmax}}\, p\left(X^{(n)|w_i}\right), \tag{1}$$

where x_t is the t-th sample in the set of n observations $X^{(n)}$.

The *max rule* chooses the identity with the highest estimated likelihood of a single face image, while the *mean* rule prefers the identity that gives the highest mean likelihood over the input images. The *majority rule*, which is an instance of a voting strategy, observes classification decisions made on all of the input images separately, and picks the label assigned to the largest number of images. Though lacking clear statistical interpretation, the same combination rules can be applied to scores instead of likelihood values (e.g., taking the product or mean of the distances in feature space).

2.2 Early integration and the MSM algorithm

The above combination rules are generally "late" integration strategies, in that they combine likelihoods rather than observations. An alternative approach consists of combining the input images themselves and mapping the whole sequence to a feature space in which the classification is performed. The simplest example of such a technique is classification of $X^{(n)}$ based on the mean image $\bar{x}$. More sophisticated approaches capture higher-order moments of the distribution. In particular, the Mutual Subspace Method (MSM) of Yamaguchi *et al* [18] is noteworthy.

In MSM, a test set of images is represented by the linear subspace spanned by the principal components of the data. This subspace is then compared to each of the subspaces constructed for the training data with dissimilarity measured by the minimal principal angle between the subspaces[1]. This approach works even with coarsely sampled sequences, or with general unordered sets of observations. However, it does not consider the entire probabilistic model of face variation, since the eigenvalues corresponding to the principal components, as well as the means of the samples, are disregarded in the comparison.

The schematic examples in Figure 1 illustrate the shortcomings of ignoring the eigenvalues and means. In Figure 1(a) the three 2D ellipses correspond to the two principal components of three data sets of 300 points obtained from 3D Gaussian distributions p_0 (solid line), p_1 (dotted) and p_2 (dashed) with the same mean but different covariance matrices. In fact p_0 is the same as p_1 but with noise that slightly rotates the ellipse plane. In terms of distribution divergence – either K-L or L_2 – p_0 is closer to p_1 than to p_2. However, the Mutual Subspace approach fails to recognize this, since the 2D principal subspaces of p_1 and p_2 are identical.

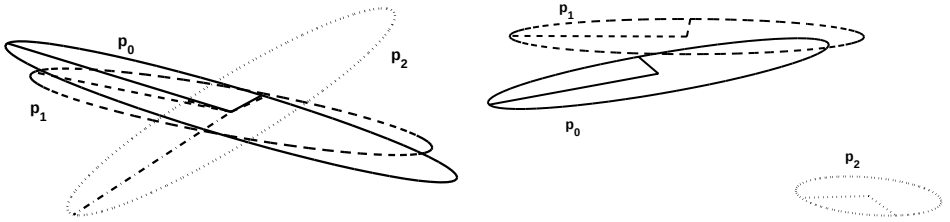
The shortcomings of the subspace angle matching are even clearer in Figure 1(b). Here the two principal components of p_0 and p_2 again lie in the same subspace, while the principal subspace of p_1 is slightly rotated. In addition, in this case important information about similarity of p_0 and p_1 is contained in the positions of the means, disregarded by the MSM.

The MSM approach has the desirable feature that it builds a compact model of the distribution of observations. However, it ignores important statistical characteristics of the data and thus, as we show in Section 4, its decisions may be statistically sub-optimal. In this paper we develop an alternative approach, which takes into account both the means and the covariances of the data and is grounded in the statistical approach to classification.

3 Statistical recognition framework

We assume that images of the kth person's face are distributed according to an underlying probability rule p_k, and that the input face images are distributed

[1] the minimal angle between any two vectors in the subspaces



(a) p_1 and p_2 have the same mean (b) p_1 and p_2 have different means

Fig. 1. Illustration of the difference between the Mutual Subspace Method and distribution divergence method. The 2D ellipses, embedded in 3D, correspond to the first 2 principal components of the estimated distributions. By the minimal principal angle measure, p_0 (solid lines) is closer to p_2 (dotted), while in terms of distribution similarity – KL-divergence or L_2-norm – p_1 (dashed) must be chosen.

according to some p_0. The task of the recognition system is then to find the class label k^* satisfying

$$k^* = \operatorname*{argmax}_k \Pr(p_0 = p_k), \tag{2}$$

subject to $\Pr(p_0 = p_{k^*}) \geq 1 - \delta$ for a given confidence threshold δ, or to conclude the lack of such a class (i.e. to reject the input). Note that score-producing classifiers, which choose the identity that maximizes a score function and not a probability, must effectively assume that the posterior class probability of the identities is monotonic with the score function. Setting the minimal threshold on a score sufficient to make a decision is equivalent to setting δ.

In this paper, we do not deal with the rejection mechanisms, and instead assume that p_0 is in fact equal to some p_k in the database of K subjects. Therefore, given a set of images distributed by p_0, solving (2) amounts to optimally choosing between K hypotheses of the form which in statistics is sometimes referred to as the two-sample hypothesis: given two samples (in our case, the training set of the kth subject and the test set), do these two sets come from the same distribution?

When only a single sample point from p_0 is available, it is known that the optimal hypothesis test is performed by choosing k maximizing the posterior probability of p_k given this sample. When a larger sample set is available, the optimal test becomes a comparison of posteriors of different models given this set. If all the classes have equal prior probabilities, this becomes a comparison of the likelihoods of the test set under different class models. In the following section, we discuss ways of performing this test, and its relationship to the Kullback-Leibler divergence between distributions.

In reality the distributions p_k are unknown and need to be estimated from data, as well as p_0. In this paper, we follow [12] and estimate the densities in the space of frontal face images by a multivariate Gaussian. Each subject has its own

density, which is estimated based on the training samples of that subject's face. Our algorithm, described below, does not require (nor uses) temporal adjacency of the images in the input sequence.

4 Density matching and hypothesis testing

Here we show the relationship of classical hypothesis testing to density matching via K-L divergence. A conclusion of the analysis, supported by subsequent experiments, is that when recognition is done by comparing sets of observations, direct estimation of KL-divergence between densities inferred from training data (i.e. the model densities) and densities inferred from samples under test is a principled alternative to standard approaches. This follows from an analysis of the K-ary hypothesis test, which can be stated as follows:

$$\begin{aligned} H_1 &: X_0^{(n)} \sim p_1(x) \\ &\vdots \\ H_K &: X_0^{(n)} \sim p_K(x) \end{aligned} \tag{3}$$

with the goal of determining which hypothesis, H_k, best explains the data, $X_0^{(n)}$. The notation $X_k^{(n)} \sim p_j(x)$ implies that the kth sample set is drawn from the jth density. In the discussion which follows $X_0^{(n)} = \left\{x_0^1, \ldots, x_0^n\right\}$ indicates the sample set under test and $p_0(x)$ the density from which it is drawn while $X_k^{(n)}$ $(k = 1 \ldots K)$ indicates the kth sample set and $p_k(x)$ the model density inferred from that sample set.

Although not commonly used, the K-ary hypothesis test:

$$\begin{aligned} H_1 &: X_1^{(n)} \sim p_0(x) \\ &\vdots \\ H_K &: X_K^{(n)} \sim p_0(x) \end{aligned} \tag{4}$$

is, under mild conditions [10], equivalent to 3. In 3 we quantify and compare the ability of our *inferred* model densities to explain the samples under test, while in equation 4 we infer the density of our test samples and then quantify and compare its ability to explain our model samples.

Assuming that the x_i are *i.i.d.* (independent and indentically distributed) and the classes have equal prior probability, it is well known, by the Neyman-Pearson lemma (e.g. see [14]), that an optimal statistic for choosing one hypothesis over the other is the log-likelihood function; that is,

$$H_k = \underset{k}{\operatorname{argmax}} \sum_i \log\left(p_k\left(x_0^i\right)\right) \tag{5}$$

for the hypothesis test of (3) and

$$H_k = \underset{k}{\operatorname{argmax}} \sum_i \log\left(p_0\left(x_k^i\right)\right) \tag{6}$$

for hypothesis test of (4). Note the implication that *all* samples are used in determination of the best hypothesis.

Under the *i.i.d.* assumption it can be shown that the log-likelihood of samples drawn from $p_l(x)$ under model distribution $p_k(x)$ has the following expected (with respect to $p_l(x)$) and asymptotic behavior [4]:

$$E_{p_l}\left\{\frac{1}{N}\sum_{i=1}^{N}\log\left(p_k\left(x_l^i\right)\right)\right\} = -\left(H(p_l(x)) + D\left(p_l(x)\,||p_k(x)\right)\right) \tag{7}$$

$$= \lim_{N\to\infty}\left(\frac{1}{N}\sum_{i=1}^{N}\log\left(p_k\left(x_l^i\right)\right)\right) \tag{8}$$

where $D(p_l||p_k)$ is the well-known asymmetric Kullback-Leibler (K-L) divergence [10] defined as

$$D(p||q) = \int p(x)\log\left(\frac{p(x)}{q(x)}\right)\mathrm{d}x\ . \tag{9}$$

It can be shown that $D(p_l||p_k) \geq 0$, with equality only when $p_l(x) = p_k(x)$.

The K-L divergence quantifies the ability of one density to explain samples formalized by the information theoretic notion of *relative entropy*. In the context of K-ary hypothesis testing we see that asymptotically (as we gather more samples) and in expectation (given a finite sample set) that we choose the model density, $p_k(x)$, which is closest to the sample density, $p_l(x)$, in K-L divergence sense.

In this light we might recast the K-ary hypothesis test as one of trying to estimate the K-L divergence between the distribution underlying our training *data* and the distribution underlying the sample data under test. The difference between the two hypothesis tests is in the *direction* one computes the divergence. The hypothesis selection rules of 5 and 6 become

$$H_k = \underset{k}{\operatorname{argmax}} \quad -D(p_0||p_k) \tag{10}$$

$$H_k = \underset{k}{\operatorname{argmax}} \quad -\left(H(p_k) + D(p_k||p_0)\right) \tag{11}$$

respectively. Consequently, we will investigate several methods by which one can approximate the K-ary hypothesis test via direct computation/estimation of K-L divergence. Depending on the assumptions made we unavoidably introduce errors into the distribution estimates which in turn introduce errors into our estimate of the K-L divergence.

In the first, and most commonly used, method we assume a parametric density form (Gaussian) with parameters estimated from the training data. The log-likelihood of the samples under test are computed using equation 5 and plugging in the estimated Gaussian densities. This is equivalent to first finding the Gaussian densities which are closest to the "true" model densities in the K-L sense and then second, of the Gaussian distributions, finding the one which is closest in the K-L sense to the distribution of the sample under test.

Secondly, we estimate the parameters of the distribution of our test sample and subsequently evaluate the log likelihood of our training data using equation 6. This is equivalent to first finding the Guassian distribution which is closest to the "true" test distribution and then, of the training distributions, selecting that which is closest to the estimated testing density in the K-L sense.

Finally, using a recent result [19], we estimate Gaussian distribution parameters for both training and test data and compute K-L divergences analytically.

5 Classification based on Kullback-Leibler divergence

In this section we define the proposed classification scheme based on the KL-divergence between the estimated model and input probability densities, and present the computation details of our scheme. The first phase of our algorithm consists of modeling the data distribution as a single multivariate Gaussian constructed from two factors – the density in the principal subspace and the isotropic "noise" component in the orthogonal subspace. This is a fairly standard procedure. The second, novel phase is the closed-form evaluation of the KL-divergence between the models estimated for input set and the ones estimated for each subject from the training data. The resulting values can be treated as a score for classification or rejection decisions.

5.1 Computation of $D_{KL}\left(p_k \| p_0\right)$

In the general case of two multivariate distributions, evaluating $D_{KL}\left(p_k \| p_0\right)$ is a difficult and computationally expensive task, especially for high dimensions, and is typically performed by means of numeric integration, and computationally expensive, especially for high dimensions. However, a recent result [19] for the case of two normal distributions p_k and p_0 provides a closed form expression:

$$D_{KL}\left(p_0 \| p_k\right) = \frac{1}{2}\log\left(\frac{|\mathbf{\Sigma}_k|}{|\mathbf{\Sigma}_0|}\right) + \frac{1}{2}\operatorname{Tr}\left(\mathbf{\Sigma}_0\mathbf{\Sigma}_k^{-1} + \mathbf{\Sigma}_k^{-1}(\bar{x}_k - \bar{x}_0)(\bar{x}_k - \bar{x}_0)^T\right) - \frac{d}{2}, \tag{12}$$

where d is the dimensionality of the data (number of pixels in the images), $\bar{x}_k$ and $\bar{x}_0$ are the means of the training set for the kth subject and of the input set, respectively, and $\mathbf{\Sigma}_k$ and $\mathbf{\Sigma}_0$ are the covariance matrices of the estimated Gaussians for the kth subject and for the input set, respectively.

After estimating $\mathbf{\Sigma}_0$ and finding its eigendecomposition, an operation taking $\mathcal{O}(d^3)$, we can compute the determinant in (12) in linear time, by taking the product of the eigenvalues. Calculation of the matrix products will require an additional workload of $\mathcal{O}(d^3)$. Therefore, for K subjects in the database, we need $\mathcal{O}(Kd^3)$ operations.

To compute $D_{KL}\left(p_0 \| p_k\right)$, we exchange the indices 0 and k in (12). Since KL-divergence is an asymmetric measure, we should expect the results to be different for the two "directions".

5.2 Normal approximation of face appearance densities

For the sake of completeness, we describe our calculation of $\mathbf{\Sigma}_0$ and $\mathbf{\Sigma}_k$, which is a straightforward application of PCA, as done, e.g., in [12]. Let $\mathbf{\Phi}_k$ be the orthonormal matrix of eigenvectors (columns) and $\mathbf{\Lambda}_k$ be the diagonal matrix of non-decreasing eigenvalues $\lambda_{k1} \geq \lambda_{k2} \geq \ldots \geq \lambda_{kd}$ of the auto-correlation matrix $\mathbf{S}_k = (X_k^{(n)} - \bar{x}_k)(X_k^{(n)} - \bar{x}_k)^T$ of the kth set of images. We will denote by the subscript k the components of such a decomposition for the training set of images of the kth subject, and by the subscript 0 the components of the decomposition of the test image set.

The maximum likelihood estimate of a multi-dimensional Gaussian from the data is $N(\cdot\,; \bar{x}_k, \mathbf{S}_k)$. However, following the assumption that the true dimensionality of the data (with the noise excluded) is lower than d, we choose a subset of $M_k \leq n$ eigenvectors, corresponding to the desired retained energy[2] E: $\sum_{i=1}^{M_k-1} \lambda_i < E \leq \sum_{i=1}^{M_k} \lambda_i$. The chosen eigenvectors define the principal linear subspace L_k.

Our model must still explain the variance in the complementary orthogonal subspace $L_k^{\perp} \perp L_k$. Following [12], we replace each one of the remaining diagonal elements of $\mathbf{\Lambda}_k$ by their mean ρ_k. This solution, which estimates the distribution in $L_k^{\perp}$ as isotropic Gaussian noise with variance ρ_k, can be shown to minimize the KL-divergence from the true distribution, if the Gaussianity assumption is correct. The estimated density of the faces in the k-th class is then the product of the Gaussian densities in the two subspaces, and can be written in the full d-dimensional space as

$$p_k(x) = N\left(x; \bar{x}_k, \mathbf{\Sigma}_k\right), \tag{13}$$

where $\bar{x}_k$ is the mean of the images in k-th class, and

$$\mathbf{\Sigma}_k = \mathbf{\Phi}_k \mathbf{\Lambda}_k^E \mathbf{\Phi}_k^T, \quad \mathbf{\Lambda}_k^E = \begin{pmatrix} \lambda_{k1} & 0 & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & . & \lambda_{kM_k} & 0 & . & 0 \\ 0 & . & 0 & \rho_k & 0 & . \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & \ldots & \ldots & \ldots & \rho_k \end{pmatrix}, \quad \rho_k = \frac{1}{d - M_k} \sum_{i=M_k+1}^{d} \lambda_{ki}$$

Similarly, the estimated density of the observed face sequence is

$$p_0(x) = N\left(x; \bar{x}_0, \mathbf{\Sigma}_0\right). \tag{14}$$

It is important to point out that $\bar{x}_k$ and $\mathbf{\Sigma}_k$ need to be computed only once for each subject, at the time the subject is entered into the database.

6 Experiments

In order to evaluate the relative performance of the proposed algorithm, we compared it to several other integration algorithms mentioned above. The comparisons were made on two data sets, described below, both containing frontal

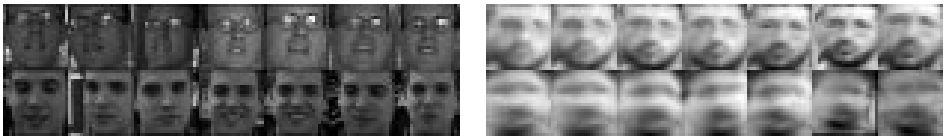
[2] alternatively one can set M_k directly; while both techniques are heuristic, we found the latter to be less robust with respect to discovering the true dimensionality

face images. One data set was obtained with a single camera using a face detector to collect images, while the other used a recently proposed virtual view rendering technique to generate frontal pose images from a set of arbitrary views.

In our experiments we compared the performance of the following algorithms:

- Our distribution comparison algorithm based on KL-divergence
- MSM
- Comparison by the mean log-likelihood of the set
- Comparison by the log-likelihood of the mean face
- Max / min / mean combination rules for likelihoods of individual images
- Max / min / mean combination rules for classifier scores of individual images (distances from the principal subspace).
- Majority vote, both by likelihoods and by scores of individual images

Due to space constraints, and to keep the graphs scaled conveniently, we report only the results on those methods that achieved reasonably low classification error for some energy level. Methods not reported were performance significantly inferior to the reported ones.



(a) Conversation video (b) Visual Hull generated face views

Fig. 2. Representative examples from the data used in the experiments. Note thethe pose and lighting variation in (a), and synthetic rendering artifacts in (b)

6.1 "Conversation" face sequences

To evaluate the discussed methods on conventional data obtained with a single monocular camera, we experimented with a data set[3] containing video clips of 29 subjects filmed during a conversation; we refer to this data set as the Conversation data. Each person is represented by one clip captured with a CCD camera, at 30 fps. The faces were detected using a fast face detection algorithm [16], and resized to 22×22 pixel gray level images. False positives were manually removed. In all of the images, the pose is frontal within about 15 degrees, with an even smaller range of tilt and yaw. The first 500 frames for each person were used as a training set, and the rest were used for testing. To estimate the behavior

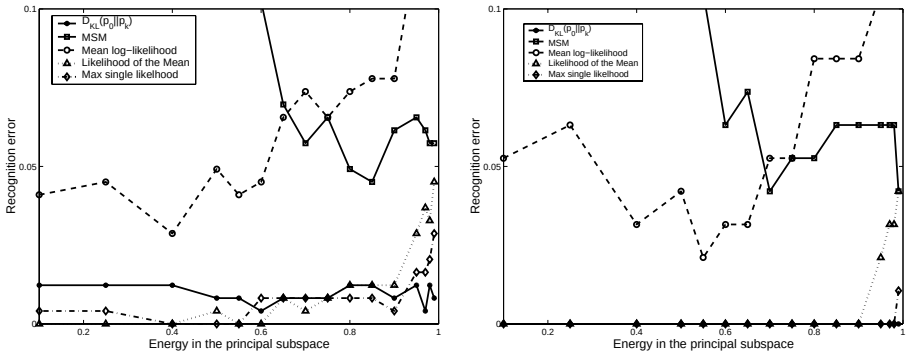
[3] courtesy of Mitsubishi Electric Research Lab

of the discussed algorithms for different input sizes, we partitioned the test images into sets of 60, 150 and 500 frames (2, 5 and 17 seconds) - 240, 95 and 25 sets, respectively. The examples shown in Figure 2(a) are representative of the amount of variation in the data.

We decided to perform the recognition task on low-resolution images (22×22) for two reasons. One is the interest to compare performance to that on virtually rendered images, described in the following section, which are small by necessity. The second is the reduction in computation and storage space.

Figure 3 shows the results for 60 and 150 frames. In both cases, $D_{KL}\left(p_k \| p_0\right)$ achieved performance statistically indistinguishable from that of the other top-performing algorithms. For sets of 60 images its error rate was 0.04%, while for sets of 150 frames its performance (and also that of a few other algorithms) was perfect. The optimal energy cut-off point for most algorithms seems to be around 0.7, but some algorithms, including ours, show robustness around this point, staying at low error for a much wider range of E. In the absence of a clear principled way of choosing the dimensionality of the principal subspace, this is an encouraging quality.

In our experiments with sets of 500 frames, the difference between many of the algorithms vanishes as they achieve zero error for some PCA dimension. We did observe, however, more robust performance with respect to PCA dimensionality, with the KL-divergence, mean face likelihood, and max rule on the likelihoods.

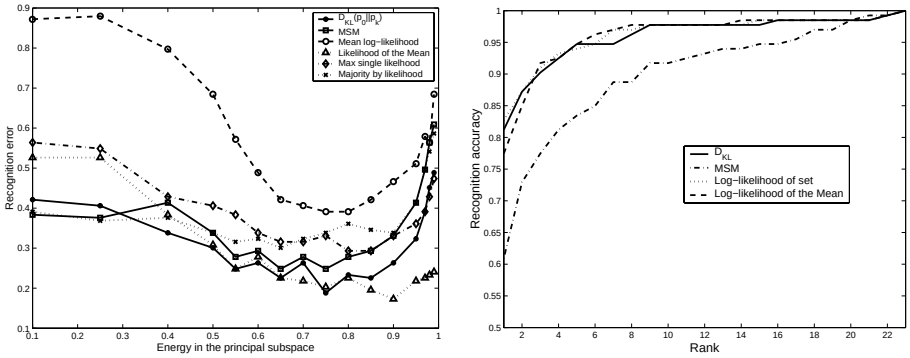


(a) 60 frames (b) 150 frames

Fig. 3. Experiments with Conversation data, sequences of (a) 60 and (b) 150 frames. Note that for 150 frames, three algorithm achieve zero error rate.

6.2 View-independent recognition with rendered images

We have previously developed methodology for view-independent multi-modal recognition using multiple synchronized views. In this framework, views of the subject are rendered from desired viewpoints using a fast implementation of image-based Visual Hulls (VH). The viewpoints are set based on the subject's pose, estimated from the trajectory, and according to the mode of operation of the relevant classifiers. The identity is then derived from combining an eigenspace-based face recognizer, operating on nearly frontal views of the face, and a gait recognizer, which extracts geometric features form a sequence of profile silhouettes [15].



(a) Performance of set recognition algorithms for different PCA dimensions

(b) Performance as a function of rank-threshold

Fig. 4. Experiments with Visual Hull- generated data, sequences of about 60 frames

The input images to the face classifier are rendered using a virtual camera placed to obtain a frontal view of the user. Thus the face image is assumed to be in roughly canonical scale and orientation, to this extent solving the pose variation problem. However, the problem of changing expression, hair style etc. remains. In addition, images are of low resolution (22×22 in our implementation) due to the typical distance of about 10 meters between the subjects and the physical cameras. Some examples of images used for training and as input to the face recognizer in our VH-based system are given in Figure 2(b). In general, recognition performance on this data set is lower than it would be on typical data. Error rate of 20% was reported in [15] for face recognition on a larger data set, that included very short input sequences.

It is also worth mentioning that the frontal face view is sought within a small spatial angle of the initial estimate of the face orientation, and often a face is detected in a number of virtual views that are nearly frontal. Therefore, the

number of available virtual face images for t time frames is typically larger than t, sometimes by an order of magnitude.

The results shown in Figure 4 were computed by leave-one-out cross validation. Each one of the 133 sets of at least 60 synthetic faces of one of the 23 subjects was used as the input, while the rest of the data was used for training. All of the classification algorithms were evaluated for different energy thresholds E varying from 0.1 (for which typically a single principal component is chosen) to .99 (typically around 100 components). As in the experiments with the Conversation data, we did not address the issue of choosing the right E, but it can be seen from the plots that in all data sets, the E producing the best results is in the range of 0.75–0.9. And again the behavior of some algorithms, including ours, is notably more robust than the others, with respect to the choice of E.

For this dataset, $D_{KL}(p_0 \| p_k)$ achieves best performance for $E = 0.75$ at misclassification rate of 18%, compared to 24% for the log-likelihood of the input set computed as in (5). For most values of E the error rate associated with $D_{KL}(p_0 \| p_k)$ is lower than the other methods. To quantify the robustness of the method, Figure 4(b) shows the classification performance as a function of rank-threshold. That is, the vertical axis corresponds to the percentage of the correct identities within the first n, as ranked by different measures. The robust performance of our algorithm (solid line) is similar to that of the mean face likelihood,

7 Discussion, conclusions and future work

Our algorithm for recognition from a set of observations is based on classifying a model built from the set, rather than classifying the individual observations and combining the results. This approach, motivated by casting the classification of sets as a statistical hypothesis testing problem, while not uncommon in the statistical community, to the best of our knowledge has not been used in vision. Our experimental results are in accordance with the theoretical analysis, and support using statistically-motivated density matching for classification. An additional potential benefit of our algorithm, though not fully realized in our current implementation, is in further reducing the computational costs of recognition by using sequential PCA and more sophisticated matrix manipulation algorithms.

We intend to continue the experiments while extending the range of the data to a larger number of classes and greater variability. For such data, the assumption that the underlying distributions are Gaussian, which allows simple computation, may not be true. We are currently investigating using alternative statistical models.

Our current integration strategy ignores the dynamics of the sequence. In fact, it classifies sets rather than sequences of images, and therefore does not assume meaningful temporal constraints between the images. We expect that including the dynamics of the face appearance, when available, into the algorithm would improve the classification.

Another interesting direction is to extend the distribution matching approach to features other than principal components of images. For instance, one may estimate the distributions of salient features, using nostril, eye, mouth etc. detectors. Alternatively, an Active Appearance Model may be fit to each image in the training and input sets, and the distributions of the computed feature values may be compared.

Finally, the general approach introduced in this paper can be applied to recognition not only of faces, but of any objects with variation in appearance across images.

References

1. Peter N. Belhumeur, Joao P. Hespanha, and David J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, July 1997.
2. Zoran Biuk and Sven Loncaric. Face Recognition from Multi-Pose Image Sequence. In *Proceedings of 2nd Int'l Symposium on Image and Signal Processing and Analysis*, pages 319–324, Pula, Croatia, 2001.
3. R. Brunelli and T. Poggio. Face recognition: features vs. templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(10):1042–1052, 1993.
4. Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., New York, 1991.
5. Trevor Darrell, David Demirdjian, Neal Checka, and Pedro Felzenswalb. Plan-view trajectory estimation with dense stereo background models. In *Proceedings of the International Conference on Computer Vision*, Vancouver, BC, July 2001.
6. G.J. Edwards, C.J. Taylor, and T.F. Cootes. Improving Identification Performance by Integrating Evidence from Sequences. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 486–491, 1999.
7. S. Gong, A. Psarrou, I. Katsoulis, and P. Palavouzis. Tracking and Recognition of Face Sequences. In *European Workshop on Combined Real and Synthetic Image Processing for Broadcast and Video Production*, pages 96–112, Hamburg, Germany, 1994.
8. Ismail Haritaoglu, David Harwood, and Larry S. Davis. W4: Real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), August 2000.
9. Josef Kittler, Mohamad Hatef, Robert P.W. Duin, and Jiri Matas. On Combining Classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, March 1998.
10. Solomon Kullback. *Information Theory and Statistics*. John Wiley and Sons, New York, 1959.
11. Baback Moghaddam, Tony Jebara, and Alex Pentland. Bayesian face recognition. *Pattern Recognition*, 33:1771–1782, 2000.
12. Baback Moghaddam and Alex Pentland. Probabilistic visual learning for object representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):696–710, 1997.
13. Penio S. Penev and J. J. Atick. Local feature analysis: A general statistical theory for object representation. *Network: Computation in Neural Systems*, 7(3):477–500, 1996.

14. Louis L. Scharf. *Statistical signal process: detection, estimation, and time series analysis*. Addison-Wesley Publishing Company, New York, 1990.
15. Gregory Shakhnarovich, Lily Lee, and Trevor Darrell. Integrated Face and Gait Recognition From Multiple Views. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, Lihue, HI, December 2001.
16. Paul A. Viola and Michael J. Jones. Robust real-time object detection. Technical report, COMPAQ Cambridge Research Laboratory, Cambridge, MA, February 2001.
17. L. Wiskott, J. M. Fellous, N. Kruger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. In *Proceedings of International Conference on Image Processing*, pages 456–463, Heidelberg, 1997. Springer-Verlag.
18. Osamu Yamaguchi, Kazuhiro Fukui, and Ken-ichi Maeda. Face recognition using temporal image sequence. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 318–323, Nara, Japan, 1998.
19. Shintaro Yoshizawa and Kunio Tanabe. Dual differential geometry associated with the Kullback-Leibler information on the Gaussian distributions and its 2-parameter deformations. *SUT Journal of Mathematics*, 35(1):113–137, 1999.