Probabilistic Linear Discriminant Analysis for Inferences About Identity

Simon J.D. Prince
Department of Computer Science
University College London
s.prince@cs.ucl.ac.uk

James H. Elder
Center for Vision Research
York University
jelder@yorku.ca

Abstract

Many current face recognition algorithms perform badly when the lighting or pose of the probe and gallery images differ. In this paper we present a novel algorithm designed for these conditions. We describe face data as resulting from a generative model which incorporates both within-individual and between-individual variation. In recognition we calculate the likelihood that the differences between face images are entirely due to within-individual variability. We extend this to the non-linear case where an arbitrary face manifold can be described and noise is position-dependent. We also develop a "tied" version of the algorithm that allows explicit comparison across quite different viewing conditions. We demonstrate that our model produces state of the art results for (i) frontal face recognition (ii) face recognition under varying pose.

1. Introduction

In current face recognition systems, the subject is required to cooperate with the system: they must stand in a certain place, face the camera and maintain a neutral expression. Under these *controlled* imaging conditions, recognition algorithms perform well. One of the greatest remaining challenges is to recognize faces in *uncontrolled conditions*. Now the subject may be entirely unaware of the system, and consequently the position, pose, illumination and expression of their face exhibit considerable variation. The ability to cope with this variation would permit recognition from surveillance footage, face search in archived images and more transparent access control. Unfortunately, in such uncontrolled conditions, most current commercial and academic face recognition systems flounder.

Many face recognition algorithms use a "distance-based" approach. (e.g. [18]). The probe and gallery images are linearly projected to a lower dimensional representation to form probe and gallery feature vectors. A match is assigned based on the distances between these vectors. A notable sub-category of these methods consists of approaches based on linear discriminant analysis (LDA). The Fisherfaces algorithm [1] projected face data to a space where the ratio of between-individual variation to within-individual variation was maximized. Fisherfaces is limited to directions in which at least some within-individual variance has been observed (the small-sample problem). The null-space LDA approach [5] exploited the signal in the remaining subspace. The Dual-Space LDA approach [19] combined these two approaches. These methods produce high-quality results but cannot always cope with large pose and illumination changes. In these cases, most of the signal lies in part of the subspace where the noise is also great. These directions are downweighted or discarded by linear LDA approaches.

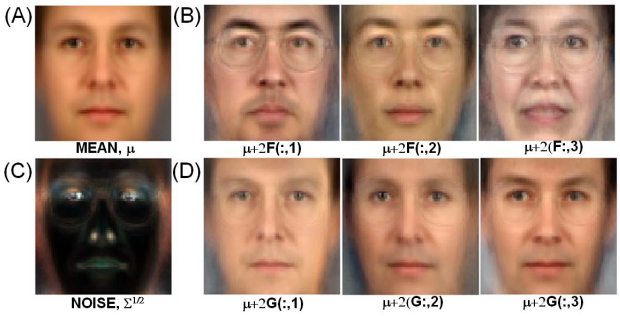


Figure 1. Components of PLDA Model. (A) Mean face (B) Three directions in between-individual subspace. Each image looks like a different person. (C) Per-pixel noise covariance (D) Three directions in within-individual subspace. Each images looks like the same person under minor pose and lighting changes.

Many alternative approaches have been suggested for recognition with variable pose and illumination. Important categories include algorithms which (i) require more than one input image of each face [8] (ii) create a 3D model from the 2D image and estimate pose and lighting explicitly [3, 2] and (iii) learn a statistical relation between faces viewed under different conditions [9, 12].

In recent work, Prince and Elder [14] proposed an novel method for recognition across large pose changes. They proposed a generative model to explain variation in the face

978-1-4244-1631-8/07/$25.00 ©2007 IEEE

data. Some of the variables in the model represented identity and others represented pose. Rather than basing recognition on distance comparisons, they calculated the likelihood that the underlying identity component was the same, regardless of the pose value. This method produced good results despite only using an impoverished per-pixel model of within-individual noise.

In this paper, we develop a probabilistic approach similar to that of [14, 15] to support three main contributions: (i) In Section 2 we introduce a probabilistic version of Fisherfaces [1], which we term probabilistic LDA (or PLDA). Similarly to [14] we explain the observed face images as the result of a generative model. We show that this approach sidesteps the small sample problem and produces superior results for frontal faces. (ii) In Section 3 we introduce a non-linear generalization of this approach. (iii) In Section 4 we introduce "Tied PLDA" which allows us to compare faces captured at very different poses.

2. Probabilistic LDA (PLDA)

Linear discriminant analysis (LDA) is a technique that models both intra-class and inter-class variance as multi-dimensional Gaussians. It seeks directions in space that have maximum discriminability and are hence most suitable for supporting class recognition tasks. In this section we present a probabilistic approach to the same problem which we term probabilistic LDA or PLDA. The relationship between PLDA and standard LDA is analagous to that between factor analysis and principal components analysis.

We assume that the training data consists of J images each of I individuals. We denote the j'th image of the i'th individual by $\mathbf{x}_{ij}$. We model data generation by the process:

$$\mathbf{x}_{ij} = \mu + \mathbf{F}\mathbf{h}_i + \mathbf{G}\mathbf{w}_{ij} + \epsilon_{ij} \tag{1}$$

This model comprises two parts: (i) the signal component $\mu + \mathbf{F}\mathbf{h}_i$ which depends only on the identity of the person but not the particular image (there is no dependence on j). This describes between-individual variation. (ii) the noise component $\mathbf{G}\mathbf{w}_{ij} + \epsilon_{ij}$ which is different for every image of the individual and represents within-individual noise.

The term μ represents the overall mean of the training dataset. The columns of the matrix $\mathbf{F}$ contain a basis for the between-individual subspace and the term $\mathbf{h}_i$ represents the position in that subspace. The matrix $\mathbf{G}$ contains a basis for the within-individual subspace and $\mathbf{w}_{ij}$ represents the position in this subspace. Remaining unexplained data variation is explained by the residual noise term ϵ_{ij} which is defined to be Gaussian with diagonal covariance Σ. The parameters $\theta = \{\mu, \mathbf{F}, \mathbf{G}, \Sigma\}$ are depicted in Figure 1.

In the parlance of factor analysis, the matrices $\mathbf{F}$ and $\mathbf{G}$ contain factors and the latent variables $\mathbf{h}_i$ and $\mathbf{w}_{ij}$ are factor loadings. For readers familiar with LDA, the columns of $\mathbf{F}$ are roughly equivalent to the eigenvectors of the between-individual covariance matrix, and the columns of $\mathbf{G}$ are roughly equivalent to the eigenvectors of the within-individual covariance matrix. The term $\mathbf{h}_i$ is particularly important as this represents the identity of individual i. We term this a *latent identity variable*: in recognition we will consider the likelihood that two face images were generated from the same underlying $\mathbf{h}_i$.

More formally, we can describe the model in Equation 1 in terms of conditional probabilities:

$$Pr(\mathbf{x}_{ij}|\mathbf{h}_i, \mathbf{w}_{ij}, \theta) = \mathcal{G}_{\mathbf{x}}\left[\mu + \mathbf{F}\mathbf{h}_i + \mathbf{G}\mathbf{w}_{ij}, \Sigma\right] \tag{2}$$

$$Pr(\mathbf{h}_i) = \mathcal{G}_{\mathbf{h}}\left[0, \mathbf{I}\right] \tag{3}$$

$$Pr(\mathbf{w}_{ij}) = \mathcal{G}_{\mathbf{w}}\left[0, \mathbf{I}\right] \tag{4}$$

where $\mathcal{G}_{\mathbf{a}}\left[\mathbf{b}, \mathbf{C}\right]$ represents a Gaussian in $\mathbf{a}$ with mean $\mathbf{b}$ and covariance $\mathbf{C}$. In Equations 3 and 4 we have defined simple priors on the latent variables $\mathbf{h}_i$ and $\mathbf{w}_{ij}$.

There are two phases to using this model. In the *training* phase, we aim to learn the parameters $\theta = \{\mu, \mathbf{F}, \mathbf{G}, \Sigma\}$ from a set of training data $\mathbf{x}_{ij}$. These remain fixed during the *recognition phase* in which we make inferences about whether faces match. We treat each of these in turn.

2.1. Training

We aim to take a set of data points $\mathbf{x}_{ij}$, and find the parameters, $\theta = \{\mu, \mathbf{F}, \mathbf{G}, \Sigma\}$ under which the data is most likely. This would be easy if we knew the values of the latent variables $\mathbf{h}_i$ and $\mathbf{w}_{ij}$. Likewise it would be easy to estimate $\mathbf{h}_i$ and $\mathbf{w}_{ij}$ given θ. Unfortunately, none of the terms on the right hand side of Equation 1 are known.

Luckily, there is a well-known solution to this chicken-and-egg problem. The Expectation Maximization (EM) algorithm [6] alternately estimates the two sets of parameters in such a way that the likelihood is guaranteed to increase at each iteration. More specifically, in the Expectation- or E-Step, we calculate a full posterior distribution over the latent variables $\mathbf{h}_i$ and $\mathbf{w}_{ij}$ for fixed parameter values. In the Maximization- or M-Step, we optimize point estimates of the parameters $\theta = \{\mu, \mathbf{F}, \mathbf{G}, \Sigma\}$. The details of this are rather involved and are presented in Appendix A.

2.2. Recognition

In recognition, we compare the likelihood of the data under R different models $\mathcal{M}_{1...R}$. We define a model $\mathcal{M}$ as representing a relationship between the underlying identity variables, $\mathbf{h}_i$ and the data (see Figure 2). If two or more faces belong to the same person, then they must have the same identity variable $\mathbf{h}_i$. If two faces belong to different people they will have different identity variables. For the q'th model we calculate a likelihood term $Pr(\mathbf{X}|\mathcal{M}_q)$

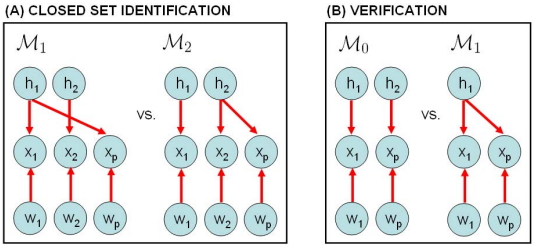


Figure 2. Recognition by comparing the likelihood of the data under different models. Each model represents a different relationship between the hidden identity variables $\mathbf{h}$ and observations $\mathbf{x}$. (A) Face identification with gallery of two faces. In Model $\mathcal{M}_1$ the probe $\mathbf{x}_p$ matches gallery face x_1. In model $\mathcal{M}_2$ the probe x_p matches $\mathbf{x}_2$. (B) Face verification. In model $\mathcal{M}_0$ the faces x_p and x_1 do not match. In model $\mathcal{M}_1$ they match.

where $\mathbf{X}$ is all of the observed data. We calculate a posterior probability for which model is correct using Bayes' rule:

$$Pr(\mathcal{M}_q|\mathbf{x}) = \frac{Pr(\mathbf{x}|\mathcal{M}_q)Pr(\mathcal{M}_q)}{\sum_{r=0}^{R} Pr(\mathbf{x}|\mathcal{M}_r)Pr(\mathcal{M}_r)} \tag{5}$$

To make these ideas concrete, we will consider the case of face identification with two gallery faces $\mathbf{x}_1$ and $\mathbf{x}_2$ and a probe face $\mathbf{x}_p$. In this case, there are two models (see Figure 2A): in model $\mathcal{M}_1$ the probe image $\mathbf{x}_p$ matches gallery image $\mathbf{x}_1$ and hence shares the latent identity variable $\mathbf{h}_1$. Gallery image $\mathbf{x}_2$ has its own identity variable. In model $\mathcal{M}_2$ the probe image $\mathbf{x}_p$ matches $\mathbf{x}_2$ and now these images share the identity variable $\mathbf{h}_2$. By way of example we will demonstrate how to calculate the likelihood of the data under model $\mathcal{M}_1$, which can be broken down into:

$$Pr(\mathbf{x}_{1,2,p}|\mathcal{M}_1) = Pr(\mathbf{x}_{1,p}|\mathcal{M}_1)Pr(\mathbf{x}_2|\mathcal{M}_1) \tag{6}$$

since the random variables associated with $\mathbf{x}_{1,p}$ and $\mathbf{x}_2$ are independent (as evidenced by that lack of connections in Figure 2A). We treat each term separately. In each case, we aim to calculate the likelihood of the observed data. Unfortunately, we don't know the values of the associated latent variables $\mathbf{h}$ and $\mathbf{w}$. We proceed by writing the joint likelihood of all observed and hidden variables, and then marginalizing over the unknown hidden variables. Hence, the first term in Equation 6 becomes:

$$Pr(\mathbf{x}_{1,p}|\mathcal{M}_1) = \int\int\int Pr(\mathbf{x}_1, \mathbf{x}_p, \mathbf{h}_1, \mathbf{w}_1, \mathbf{w}_p) d\mathbf{h}_1 d\mathbf{w}_1 d\mathbf{w}_p \tag{7}$$

This can be rewritten as:

$$Pr(\mathbf{x}_{1,p}|\mathcal{M}_1) = \int \left[\int Pr(\mathbf{x}_1|\mathbf{h}_1, \mathbf{w}_1) Pr(\mathbf{w}_1) d\mathbf{w}_1 \int Pr(\mathbf{x}_p|\mathbf{h}_1, \mathbf{w}_p) Pr(\mathbf{w}_p) d\mathbf{w}_p \right] . Pr(\mathbf{h}_1) d\mathbf{h}_1 \tag{8}$$

where we have rewritten the joint probability in terms of conditional probabilities in the second line. Likewise the second term in Equation 6 becomes:

$$Pr(\mathbf{x}_2|\mathcal{M}_1) = \int\int Pr(\mathbf{x}_2|\mathbf{h}_2, \mathbf{w}_2) Pr(\mathbf{w}_2) d\mathbf{w}_2 Pr(\mathbf{h}_2) d\mathbf{h}_2 \tag{9}$$

Notice that all of the conditional probabilities in these expressions were defined in the initial description of the model in Equations 2, 3 and 4. The probability of the data under model $\mathcal{M}_2$ can be decomposed in a similar way.

The process of integrating out, or marginalizing the hidden variables, $\mathbf{h}$ and $\mathbf{w}$ has the following interpretation: we are interested in finding the probability that faces have the same identity. However, we recognize that we have observed this identity under noisy conditions, and do not calculate a point estimate $\hat{\mathbf{h}}$ of identity. Rather, we calculate the probability that the two faces had the same identity, regardless of what this actual identity was. We simultaneously consider all possible instantiations of the within-individual noise.

An interesting side-effect of the marginalization is that it is valid to compare models with different numbers of identity variables $\mathbf{h}$. Consider the case of face verification (Figure 2B). We compare model $\mathcal{M}_1$ where two faces match (have the same underlying identity variable) to model $\mathcal{M}_0$ where they do not (different underlying identity variables.) This is an example of Bayesian model selection.

The PLDA model is linear with Gaussian noise, so it is possible to compute the integrals that comprise Equations 8 and 9 exactly: in general the problem is to evaluate the the likelihood that N images $\mathbf{x}_{1\ldots N}$ share the same identity variable, $\mathbf{h}$, regardless of the noise variables $\mathbf{w}_1 \ldots \mathbf{w}_N$. We can achieve this by combining the generative equations for all of these N images to form a composite system:

$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{bmatrix} = \begin{bmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{bmatrix} + \begin{bmatrix} \mathbf{F} & \mathbf{G} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{F} & \mathbf{0} & \mathbf{G} & \ldots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{F} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{G} \end{bmatrix} \begin{bmatrix} \mathbf{h} \\ \mathbf{w}_1 \\ \mathbf{w}_2 \\ \vdots \\ \mathbf{w}_N \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix} \tag{10}$$

or, giving names to these composite matrices:

$$\mathbf{x}' = \mu' + \mathbf{A}\mathbf{y} + \epsilon' \tag{11}$$

We can rewrite this compound model in terms of probabilities to give:

$$Pr(\mathbf{x}'|\mathbf{y}) = \mathcal{G}_{\mathbf{x}'}[\mathbf{A}\mathbf{y}, \Sigma'] \tag{12}$$

$$Pr(\mathbf{y}) = \mathcal{G}_{\mathbf{y}}[0, \mathbf{I}] \tag{13}$$

where

$$\Sigma' = \begin{bmatrix} \Sigma & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \Sigma & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \Sigma \end{bmatrix} \tag{14}$$

This now has the form of a standard factor analyzer, whose likelihood is well known to be:

$$Pr(\mathbf{x}_{1\dots N}) = Pr(\mathbf{x}') = \mathcal{G}_{\mathbf{x}'}\left[\mu', \mathbf{A}\mathbf{A}^T + \Sigma'\right] \tag{15}$$

In practice, the known structure of the matrix $\mathbf{A}$ can be exploited to compute this efficiently.

2.3. Datasets

Throughout this paper, we present results for the XM2VTS database using two different types of data representation. In Experiments 1,3 and 4 we use *minimally pre-processed data*. Raw pixel values form the input vector. There has been no photometric normalization. Image registration is only affine. Moreover, the training and test images are non-overlapping and the probe and gallery images are from different sessions. These characteristics mean that recognition performance will never be high, but it is easy to compare the relative inferential power of algorithms. In Experiments 2, and 5 we use *elaborately pre-processed data* to yield state of the art results and compare to published data.

Minimal Preprocessing: Each image was segmented with an iterative graph-cuts procedure. Three points were marked by hand. Faces were normalized to a standard template using an affine transform. Final size was $70 \times 70 \times 3$. The unprocessed pixels from these images were used as input to the PLDA algorithm. In each experiment we trained the system using 4 images each of the first 195 people in the database. The test set comprises 1 gallery and 1 probe image from each of the remaining 100 people. These were taken from the first and last recording sessions respectively.

Elaborate Preprocessing: Eight keypoints on each face were identified by hand. The images were registered to a standard template using a piecewise triangular warp. The final image size was 400×400. We extract a feature vector consisting of image gradients at 8 orientations and 3 scales at points in a 6×6 grid around each keypoint. A separate PLDA model was built for each keypoint. The likelihoods from the 21 submodels are assumed to be independent. Hence, we take the product to calculate the overall likelihood in Equation 5.

2.4. Experiments 1 and 2

In experiment 1 we investigate face identification using the minimally preprocessed frontal dataset. We applied 6

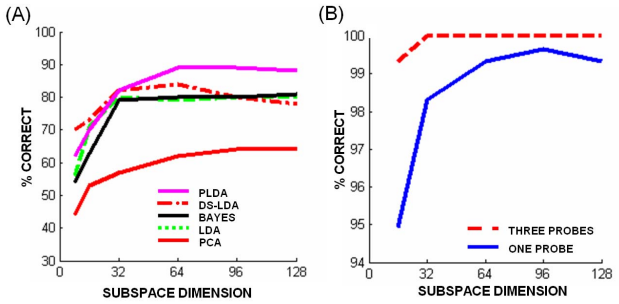


Figure 3. (A) Experiment 1 results. PLDA outperforms PCA [18], LDA [1], the Bayesian approach [13] and Dual-Space (DS) LDA [19] (B) Experiment 2 results. With one probe image, PLDA acheives a peak of 294/295 correct matches. With three probe images all images are matched correctly.

iterations of the EM algorithm, initializing the model parameters θ to random values. The results of this learning process are shown in Figure 1. Notice that as we move along the axes of the signal subspace $\mathbf{F}$ the resulting images look like different people. As we move along the axes of the noise subspace $\mathbf{G}$, the resulting images look like the same person with slightly different poses and illuminations.

On each trial, the algorithm selects the matching gallery image, by choosing the maximum a posteriori (MAP) model from Equation 5 with uniform priors. In Figure 3A we plot $\%$ correct first match results as a function of the subspace dimension: the signal and noise parameters are the only two free parameters in our model, and we set these to be identical in all experiments. We also plot results from our implementations of five other algorithms. The PLDA method outperforms all methods on this task. The closest competing method is dual-space LDA [19].

In Experiment 2 we test the same method using the elaborate pre-processing method. We employed a protocol that matches published results to facilitate quantitative comparison. We trained our system using images from the first three sessions from all 295 individuals in the XM2VTS database. We use 295 images from the fourth session to form a probe set. The gallery set comprised either (i) one image from the first session or (ii) three images of each person, one each from the first three sessions.

The $\%$ first match correct results are plotted as a function of the subspace dimension in Figure 3B. With a single probe, the peak performance is 99.4%: we only mis-classify one face. Examining this face (No. 169.4.1) reveals that the pose deviates significantly from frontal. In Section 4 we present an algorithm to cope with significant pose changes. With three probes we easily achieve 100% performance.

These results compete with the best modern algorithms. In Table 1 we show results of other algorithms tested with the same protocol. Our algorithm compares favorably, although it is unwise to draw strong conclusions where the

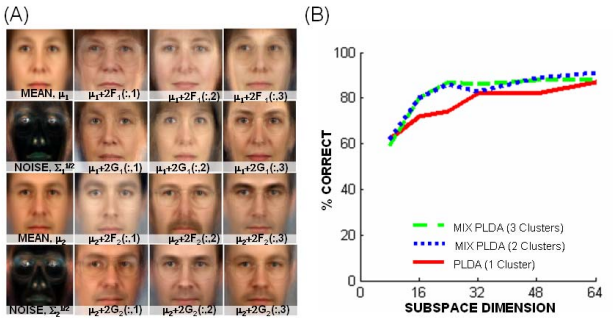


Figure 4. (A) Mixtures of PLDA model. Top two rows show elements of mixture component 1. Bottom two rows show component 2. Interestingly, the two clusters correspond to the two sexes. The mean of cluster 1 and images representing directions in the signal subspace all look like women (top row). Similarly, for cluster 2 (row three) images all look like men. As before, different positions in the within-individual subspace (2nd and 4th row) look like different images of the same person. (B) Identification results from MixPLDA model as a function of subspace dimension.

difference in performance may be only a single classified face. We believe that Figure 3A provides more information about the relative strengths of these algorithms.

METHOD	N	Error Rate
PCA [18]	1	33.9
LDA [1]	1	11.9
Bayesian [13]	1	11.5
Unified Subspace (US)[20]	1	6.8
Adaptive Clustering US SVM [11]	1	1.0
Our Approach	1	0.3
Bayesian Gabor [21]	3	2.9
Our Approach	3	0.0

Table 1 - Results for XM2VTS database with N probe images PCA, LDA, Bayesian and Unified Subspace results from [11].

It is easy to understand why our technique outperforms other LDA methods: the inclusion of the per-pixel noise term Σ means we have a more sophisticated model of within-individual variation. Our method handles the signal subspace $\mathbf{F}$ and noise subspace $\mathbf{G}$ in a unified way without the need for two separate procedures. Unlike [19] we are not required to estimate the values of unobserved eigenvalues. In addition the method has certain other advantages: it provides a posterior probability over matches and incorporating priors is straightforward. Moreover, the probabilistic formulation paves the way for the non-linear approaches presented in Sections 3 and 4.

3. Mixtures of PLDAs

In practice, it is unrealistic to assume that the face manifold is well modelled by a linear subspace. It is also unlikely that the noise distribution is identical at each point in space. In this section, we resolve these problems by describing the face manifold as a weighted additive mixture of K PLDA distributions (which we term MixPLDA).

There are now two latent identity variables associated with an individual: the scalar term c_i determines which subspace cluster the individual belongs to, and the identity vector $\mathbf{h}_i$ that determines the position within this cluster. In order for two faces to belong to the same individual *both* of these variables must match. We can write this model as:

$$\begin{aligned} Pr(\mathbf{x}_{ij}) &= \mathcal{G}_x\left[\mu_{c_i} + \mathbf{F}_{c_i}\mathbf{h}_i + \mathbf{G}_{c_i}\mathbf{w}_{ij}, \Sigma_{c_i}\right] \\ Pr(\mathbf{h}_i) &= \mathcal{G}_h\left[\mathbf{0}, \mathbf{I}\right] \\ Pr(\mathbf{w}_{ij}) &= \mathcal{G}_w\left[\mathbf{0}, \mathbf{I}\right] \\ Pr(c_i = k) &= \pi_k \qquad\qquad k = \{0 \ldots K\} \end{aligned} \tag{16}$$

All terms have the same interpretation as before, but now there are k sets of parameters $\Theta_k = \{\mu_k, \mathbf{F}_k, \mathbf{G}_k, \Sigma_k\}$. The term π_k is the prior probability of a measurement belonging to cluster k, where there are K clusters in total.

3.1. Learning and Recognition

To learn the MixPLDA model we apply the standard recipe for learning mixtures of distributions (e.g. See [7] and [4] Chap. 9). We embed the PLDA learning algorithm inside a second instance of the EM algorithm. (i) E-Step: For fixed $\mathbf{F}_{1\ldots K}, \mathbf{G}_{1\ldots K}, \Sigma_{1\ldots K}$, calculate the posterior probability $Pr(c_i = k|\mathbf{x}_{ij})$ that a given individual i belongs to the k'th cluster using the likelihood term in Equation facLike. (ii) M-Step: for each cluster k, learn the associated PLDA model using data weighted by the posterior probability of belonging to the cluster.

In recognition, we again assess the probability that faces were generated from common underlying identity variables. This now includes the choices of cluster c_i as well as the position in that cluster $\mathbf{h}_i$. Once more, each of these quantities is fundamentally uncertain so we marginalize over all possible values. The analogue of Equation 7 is:

$$Pr(\mathbf{x}_{1,p}|\mathcal{M}_1) = \tag{17}$$
$$\sum_{c_1=1}^{k} \iiint Pr(\mathbf{x}_1, \mathbf{x}_p, \mathbf{h}_1, c_1, \mathbf{w}_1, \mathbf{w}_p,) d\mathbf{h}_1 d\mathbf{w}_1 d\mathbf{w}_p$$

Note that it would not have been possible to construct this model in a conventional LDA approach. The representation of identity consists of one discrete variable c_i and one continuous variable $\mathbf{h}_i$ and distance measurements are no longer straightforward.

3.2. Experiment 3

In Experiment 3, we repeat Experiment 1 for the mixture model. We use 10 iterations of the outer loop of the EM algorithm, and update the PLDA model at each iteration with

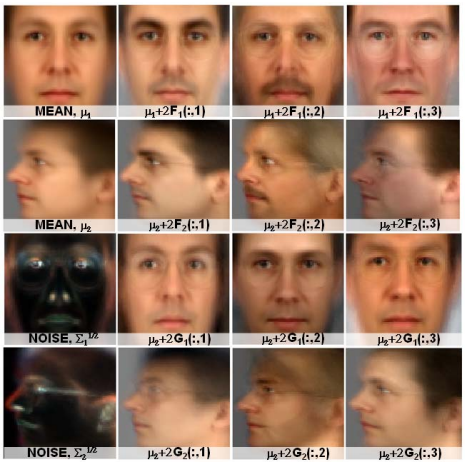


Figure 5. Tied PLDA model for face recognition across pose. The position $\mathbf{h}_i$ in the identity subspace $\mathbf{F}$ is forced to be constant for both poses: as we move along the dimensions of the signal subspace, the basis functions look like the same person, regardless of pose (top two rows). Position in the noise subspaces $\mathbf{G}$ is not tied, so these basis functions are unrelated (bottom two rows).

6 iterations as before. Examples of the learnt parameters θ for subspace dimension 8 can be found in Figure 4A for the case with K=2 clusters. Interstingly, the algorithm has organized the clusters to separate men from women.

Percent correct performance for the same dataset is plotted in Figure 4B. There is a clear improvement in performance as we move from 1 to 2 clusters, but adding a third cluster does not make much difference. However, these results should be treated with some caution: the 2 cluster mixPLDA model has twice as many parameters as the original PLDA model. In principal, it is possible for the two clusters with N/2 dimensions to exactly replicate the PLDA model with N dimensions. However, the clusters found in Figure 4A suggest that this did not happen in practice.

The case would be clearer if we could investigate higher dimensional subspaces and demonstrate a clear performance benefit from the mixture model. Unfortunately, our ability to construct the between individual subspace $\mathbf{F}$ is limited by the number of individuals in the database (195). With three clusters of 64 dimensions, this only leaves 1.01 people per dimension per cluster! Despite these concerns, we believe that the MixPLDA model is a promising method. It is fundamentally more expressive than linear methods, and retains the advantages of the probabilistic approach.

4. Tied PLDAs

Although the above methods can cope with a considerable amount of image variation, there are some cases such as large pose changes, where viewing conditions are so disparate that a more powerful technique must be applied. In "tied" models [14], two or more viewing conditions are compared by assuming that they have a common underlying variable $\mathbf{h}_i$, but different generation processes. For example, consider viewing j images each of i individuals, at k different poses. For pedagogical reasons, we will assume that the pose $\mathbf{k}$ is known for each observed datum $\mathbf{x}_{ijk}$ and there is no uncertainty over this variable. The generative model for this data is:

$$\begin{aligned} Pr(\mathbf{x}_{ijk}|\mathbf{h}_i, \mathbf{w}_{ijk}) &= \mathcal{G}_{\mathbf{x}}\left[\mu_k + \mathbf{F}_k\mathbf{h}_i + \mathbf{G}_k\mathbf{w}_{ijk} + \epsilon_{ijk}, \Sigma_k\right] \\ Pr(\mathbf{h}_i) &= \mathcal{G}_{\mathbf{h}}\left[\mathbf{0}, \mathbf{I}\right] \\ Pr(\mathbf{w}_{ijk}) &= \mathcal{G}_{\mathbf{w}}\left[\mathbf{0}, \mathbf{I}\right] \end{aligned} \tag{18}$$

Note that this model is quite different from the MixPLDA model. Both models describe the training data as a mixture of factor analyzers. However, in the mixPLDA model, the representation of identity includes the choice of cluster c_i. In the Tied PLDA model, the representation of identity $\mathbf{h}_i$ is constant (tied) *regardless* of the cluster (viewing condition). Another way to think about this is that the data is described as k clusters, but certain positions in each cluster are "identity-equivalent" to each other.

4.1. Learning and Recognition

Learning is very similar to the original PLDA model, with one major difference. In the E-Step, we calculate the posterior distribution over the latent variables given the observed data as before. However, there is now a separate M-Step for each cluster k, in which the terms $\mu_k, \mathbf{F}_k, \mathbf{G}_k, \Sigma_k$ are updated using only the data known to come from these clusters. A more detailed description of the principles behind tied models can be found in [14].

Recognition proceeds exactly as in the PLDA model, but now likelihood terms in Equation 6 are calculated by marginalizing the joint probability of data and hidden variables implicitly defined by Equations 18.

4.2. Experiments 4 & 5

We train using 195 individuals from the XM2VTS database, with 4 frontal and 4 profile faces of each individual. We test using a single frontal gallery image and right-profile probe image from the remaining 100 individuals in the database. These are taken from the 1^{st} and 4^{th} recording session respectively. Pose is always assumed to be known. For these experiments we used 10 iterations of the EM algorithm.

In Experiment 4, we use the full images with the same minimal preprocessing as in Experiment 1. In Figure 5 we present examples of the learnt basis functions $\mathbf{F}_k$ and $\mathbf{G}_k$ and noise covariance Σ. The "tied" structure is reflected in the fact that the columns of $\mathbf{F}_1$ and $\mathbf{F}_2$ look like images of the same people. In Figure 6A we plot % correct first match results as a function of the subspace dimension for

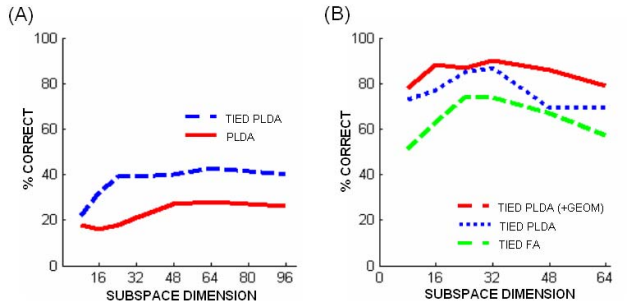


Figure 6. Results for recognition across a 90^o pose change. (A) Minimal pre-processing (B) Full model with and without geometry contribution. Results from tied factor analysis model of Prince and Elder [14] plotted for comparison.

both the tied PLDA and PLDA models. The tied PLDA model doubles performance but only from roughly 20% to 40%.

However, in Experiment 5, we apply the same preprocessing as in Experiment 2. Three of the original 8 keypoint positions are occluded in the profile model. We omit these and add one more feature on the right side of the face to compensate. Identification performance is plotted in Figure 6B. Peak performance for our algorithm is now 87%. In order to improve the results we also build a tied PLDA model of the face geometry. We align 15 keypoint positions (eyes,nose,mouth etc.) to a standard template face using a similarity transformation and a least-squares cost function. We construct a PLDA model for the (x,y) positions after this alignment. Combining this into the final likelihood calculation increases the peak performance to 92%. However, caution should be applied in interpreting this result as these keypoint positions were hand-localized.

In Table 2 we compare our method to published results. Our results compare favorably to all previous attempts at this problem, even without the contribution of geometry. The FERET database may be easier than the XM2VTS data as images were collected in the same session.

STUDY	DATABASE	POSE DIFF($^\circ$)	%
Gross [9]	FERET (100)	30($Ave.$)	75
Blanz [2]	FRVT (87)	45	86
Kim [10]	XM2VTS (125)	30	53
Prince [14]	FERET (100)	90	86
Our Approach	XM2VTS (100)	90	87

Table 2: Results for % correct face identification across large pose changes. Number of gallery images given in brackets after database name.

5. Discussion

In this paper, we have presented a probabilistic approach to LDA. The key findings are that (i) inference is more powerful in PLDA than LDA as we have a more sophisticated noise model which also involves a per-pixel noise term. (ii) the probabilistic approach allows the development of non-linear extensions that are not obvious in the standard approach. We have demonstrated that it is possible to achieve good performance with the PLDA approach, in both frontal face recognition, and recognition across large pose differences. In each case, we have endeavoured to compare to contemporary algorithms in matched conditions.

The degree of image preprocessing and feature extraction is a key determinant of performance, and it is quite possible that results might improve given different feature choices. A second possible avenue for development is suggested by the work of Wang and Tang [22] who improved LDA performance by using several LDA classifiers trained on sampled subsets of the data. In the probabilistic context, this corresponds to allowing uncertainty on the subspace matrices $\mathbf{F}$ and $\mathbf{G}$. This could be achieved using either variational Bayes or sampling methods.

Although this paper has examined face recognition, LDA is a very general technique and this method could find application in many other areas of computer vision. Although implementation is more complex, the performance is superior in every case we have tried. Moreover, the non-linear extensions provide greater expressiveness than the original LDA model. Many problems in vision are naturally expressed in generative terms as the forward problem (graphics) is well understood. Combining problem-specific generative components with abstract generative models like PLDA is a promising approach to many vision tasks.

References

[1] P.N. Belhumeur, J. Hespanha and D.J. Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection," *PAMI,* Vol. 19, pp. 711-720, 1997. 1, 2, 4, 5

[2] V. Blanz, P. Grother, P. J. Phillips and T. Vetter, "Face recognition based on frontal Views generated from non-frontal images," *CVPR,* pp. 454-461, 2005. 1, 7

[3] V. Blanz, S. Romdhani and T. Vetter, "Face identification across different poses and illumination with a 3D morphable model," *ICFGR,* pp. 202-207, 2002. 1

[4] C.M. Bishop, *Pattern Recognition and Machine Learning,* Springer, New York, 2007. 5

[5] L.F. Chen, H.Y.M. Liao, J.C.Lin, M.T. Ko, and G.J. Yu, "A new LDA-based face recognition system which can solve the small sample size problem," *Pattern Recognition*, Vol. 33, pp. 1713-1726, 2000. 1

[6] A.P. Dempster, N.M. Laird and D.B. Rubin, "Maximum likelihood for incomplete data via the EM algorithm," *Proc. Roy. Stat. Soc. B*, Vol 39, pp. 1-38, 1977. 2

[7] Z. Gharamani and G.E. Hinton, "The EM Algorithm for Mixtures of Factor Analyzers", Technical Report CRG-TR-96-1, Department of Computer Science, University of Toronto, 1996. 5

[8] A. Georghiades, P. Belhumeur and D. Kriegman, "From few to many: illumination cone models and face recognition under variable lighting and pose," *PAMI*, vol. 23, pp. 129-139, 2001. 1

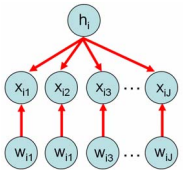
[9] R. Gross, I. Matthews and S. Baker, "Eigen light-fields and face recognition across pose," *ICAFG*, pp. 1-7, 2002. 1, 7

[10] T. Kim and J. Kittler, "Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image," *PAMI,* Vol. 27, pp. 318 - 327, 2005. 7
[11] Z. Li and X. Tang, "Bayesian face recognition using support vector machine and face clustering," *CVPR*, pp. 259-265, 2004. 5
[12] S. Lucey and T. Chen, "Learning Patch Dependencies for Improved Pose Mismatched Face Verification," *CVPR*, pp. 17-22, 2006. 1
[13] B. Moghaddam, T. Jebara and A. Pentland, "Bayesian face recognition," *Pattern Recognition,* Vol. 33, pp. 1771-1782. 4, 5
[14] S.J.D. Prince and J.H. Elder, "Tied factor analysis for face recognition across large pose changes," *BMVC*, Vol. 3, pp. 889-898, 2006. 1, 2, 6, 7
[15] S.J.D. Prince, J. Aghajanian, U. Mohammed and M. Sahani, "Latent Identity Variables: Biometric Matching without Explicit Identity Estimation," *Int'l Conf. Biometrics*, 2007. 2
[16] R. Rubin and D. Thayer, "EM algorithms for ML factor analysis," *Psychometrica,* Vol. 47, pp. 69-76, 1982. 8
[17] M.E. Tipping and C.M. Bishop, "Probabilistic principal component analysis," *J. Roy. Soc. Stat. B*, Vol. 61, pp. 611-622, 1999. 8
[18] M. Turk and A.P. Pentland, "Face recognition Using eigenfaces," *CVPR*, pp.586-591, 1991. 1, 4, 5
[19] X.Wang and X. Tang, "Dual-space linear discriminant analysis for face recognition," *CVPR,* Vol. 2, pp.564-569, 2004. 1, 4, 5
[20] X. Wang and X. Tang, "Unified Framework for Subspace Face Recognition," *PAMI*, Vol. 26, pp. 1222-1228, 2004. 5
[21] X. Wang and X. Tang, "Bayesian Face Recognition Using Gabor Features," *WBMA,* pp. 70-73, 2003. 5
[22] X. Wang and X. Tang, "Random sampling for subspace face recognition ," *IJCV*, Vol. 70, pp. 91-104, 2006. 7

Appendix 1: Learning PLDA Models

The goal of this section is to present the EM algorithm updates for learning the PLDA model described in Equation 1. The basic approach is to rewrite both E-Step and M-Step to resemble the simpler factor analysis model by assimilating terms. Updates for factor analysis are well known (see [16, 17]).

E-Step: We simultaneously estimate the joint probability distribution of all J+1 latent variables $\mathbf{h}_i, \mathbf{w}_{i1\ldots iJ}$ that pertain to each given individual i (see inset figure). We can combine together the generative equations for all of the data $\mathbf{x}_i = \{\mathbf{x}_{i1\ldots iJ}\}$ pertaining to individual i as in Equations 10 and 11, resulting in a likelihood and prior terms:



$$Pr(\mathbf{x}_i|\mathbf{y}_i\theta) = \mathcal{G}_{\mathbf{x}}\left[\mathbf{A}\mathbf{y}_i, \Sigma'\right] \quad (19)$$
$$Pr(\mathbf{y}_i) = \mathcal{G}_{\mathbf{y}}\left[0, \mathbf{I}\right] \quad (20)$$

where $\mathbf{A}$, Σ' and $\mathbf{y}_i$ are defined as in Equations 11 and 14. The model defined in Equations 19 and 20 takes the form of a factor analysis model. Applying Bayes' rule to calculate the posterior, we get:

$$Pr(\mathbf{y}_i|\mathbf{x}_i, \theta) \propto Pr(\mathbf{x}_i|\mathbf{y}_i, \theta)Pr(\mathbf{y}_i) \quad (21)$$

Since both terms on the right are Gaussian, the term on the left must be Gaussian. In fact, it can be shown that the first two moments of this Gaussian are:

$$E[\mathbf{y}_i] = (\mathbf{A}^T\Sigma'^{-1}\mathbf{A} + \mathbf{I})^{-1}\mathbf{A}^T\Sigma'^{-1}(\mathbf{x}_i - \mu') \quad (22)$$
$$E[\mathbf{y}_i\mathbf{y}_i^T] = (\mathbf{A}^T\Sigma'^{-1}\mathbf{A}^T + \mathbf{I})^{-1} + E[\mathbf{y}_i]E[\mathbf{y}_i]^T \quad (23)$$

M-Step: In the M-Step, we aim to update the values of the parameters $\theta = \{\mu, \mathbf{F}, \mathbf{G}, \mathbf{\Sigma}\}$. We rewrite Equation 1 as:

$$\mathbf{x}_{ij} = \mu + \begin{bmatrix}\mathbf{F} & \mathbf{G}\end{bmatrix}\begin{bmatrix}\mathbf{h}_i \\ \mathbf{w}_{ij}\end{bmatrix} + \epsilon_{ij} \quad (24)$$
$$= \mu + \mathbf{B}\,\mathbf{z}_{ij} + \epsilon_{ij} \quad (25)$$

where $\mathbf{B}$ is a concatenation of the two subspace matrices $\mathbf{F}$ and $\mathbf{G}$ and $\mathbf{z}_{ij}$ is a concatenation of the two factor loading vectors $\mathbf{h}_i$ and $\mathbf{w}_{ij}$. In the M-Step, we optimize:

$$Q(\theta_t, \theta_{t-1}) = \sum_{i=1}^{I}\sum_{j=1}^{J}\int Pr(\mathbf{z}_i|\mathbf{x}_{i1\ldots iJ}, \theta_{t-1})\log[Pr(\mathbf{x}_{ij}|\mathbf{z}_i)Pr(\mathbf{z}_i)]d\mathbf{z}_i \quad (26)$$

where t is the iteration index. The first log probability term in Equation 26 can be written as:

$$\log[Pr(\mathbf{x}_{ij}|\mathbf{z}_i\theta_t)] = K - 0.5\left(\log|\Sigma^{-1}| + (\mathbf{x}_{ij} - \mu - \mathbf{B}\mathbf{z}_i)^T\Sigma^{-1}(\mathbf{x}_{ij} - \mu - \mathbf{B}\mathbf{z}_i)\right) \quad (27)$$

where K is an unimportant constant. We substitute this term into Equation 26 and take derivatives with respect to $\mathbf{B}$ and Σ. The second log term in Equation 26 has no dependence on these parameters. We equate these derivatives to zero and re-arrange to provide the following update rules:

$$\mu = \frac{1}{IJ}\sum_{i,j}\mathbf{x}_{ij} \quad (28)$$
$$\mathbf{B} = \left(\sum_{i,j}(\mathbf{x}_{ij} - \mu)E[\mathbf{z}_i]^T\right)\left(\sum_{i,j}E[\mathbf{z}_i\mathbf{z}_i^T]\right)^{-1}$$
$$\mathbf{\Sigma} = \frac{1}{IJ}\sum_{i,j}\mathbf{Diag}\left[(\mathbf{x}_{ij} - \mu)(\mathbf{x}_{ij} - \mu)^T - \mathbf{B}E[\mathbf{z}_i](\mathbf{x}_{ij} - \mu)^T\right]$$

where **diag** represents the operation of retaining only the diagonal elements from a matrix. The expectation terms $E[\mathbf{z}_i]$ and $E[\mathbf{z}_i\mathbf{z}_i^T]$ can be extracted from Equations 22and 23 using the equivalence between Equations 10and 11. The updated values of $\mathbf{F}$ and $\mathbf{G}$ are retrieved from the new value of $\mathbf{B}$ using the equivalence between Equations 24 and 25.